DE GRUYTER

Marc Kupietz,
Thomas Schmidt (Hrsg.)

KORPUSLINGUISTIK

GERMANISTISCHE SPRACHWISSENSCHAFT UM 2020

DE G

Marc Kupietz und Thomas Schmidt (Hrsg.)
Korpuslinguistik

Germanistische Sprachwissenschaft um 2020

Herausgegeben von
Albrecht Plewnia und Andreas Witt

Band 5

IDS
INSTITUT FÜR
DEUTSCHE SPRACHE

Korpuslinguistik

Herausgegeben von
Marc Kupietz und Thomas Schmidt

DE GRUYTER

Die Open-Access-Publikation dieses Bandes wurde gefördert vom Institut für Deutsche Sprache, Mannheim.

ISBN 978-3-11-053674-4
e-ISBN (PDF) 978-3-11-053864-9
e-ISBN (EPUB) 978-3-11-053683-6

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

publiziert von Walter de Gruyter GmbH, Berlin/Boston
Foto Einbandabbildung: © Oliver Schonefeld, Institut für Deutsche Sprache, Mannheim
Portrait Ludwig M. Eichinger, Seite V: © David Ausserhofer, Leibniz-Gemeinschaft
Satz: Meta Systems Publishing & Printservices GmbH, Wustermark
Druck und Bindung: CPI books GmbH, Leck

www.degruyter.com

Ludwig M. Eichinger gewidmet

Vorwort

Wo steht die germanistische Sprachwissenschaft aktuell? Der vorliegende Band mit dem Titel „Korpuslinguistik" ist der fünfte Teil einer auf sechs Bände angelegten Reihe, die eine zwar nicht exhaustive, aber doch umfassende Bestandsaufnahme derjenigen Themenfelder innerhalb der germanistischen Linguistik bieten will, die im Kontext der Arbeiten des Instituts für Deutsche Sprache in den letzten Jahren für das Fach von Bedeutung waren und in den kommenden Jahren von Bedeutung sein werden (und von denen nicht wenige auch vom Institut für Deutsche Sprache bedient wurden und werden). Jeder einzelne Band behandelt ein abgeschlossenes Themengebiet und steht insofern für sich; in der Zusammenschau aller Bände ergibt sich ein Panorama der „Germanistischen Sprachwissenschaft um 2020".

Anlass des Erscheinens dieser Bände ist der Eintritt des langjährigen Direktors des Instituts für Deutsche Sprache, Ludwig M. Eichinger, in den Ruhestand. Ludwig M. Eichinger leitete das Institut von 2002 bis 2018. Seine akademische Laufbahn begann er als Wissenschaftlicher Assistent an der Universität Bayreuth; anschließend war er Heisenberg-Stipendiat an der Ludwig-Maximilians-Universität München. Ab 1990 hatte er eine Fiebiger-Professur für Deutsche Sprachwissenschaft an der Universität Passau inne, 1997 wurde er auf den Lehrstuhl für Deutsche Philologie an der Christian-Albrechts-Universität zu Kiel berufen. Mit seiner Ernennung zum Direktor des Instituts für Deutsche Sprache im Jahr 2002 wurde er auch Ordinarius für Germanistische Linguistik an der Universität Mannheim. Ludwig M. Eichinger ist Ehrendoktor der Pannonischen Universität Veszprém und der Universität Bukarest. Er ist Mitglied der Akademie der Wissenschaften und der Literatur zu Mainz und der Österreichischen Akademie der Wissenschaften; außerdem ist er Ständiger Gastprofessor an der Beijing Foreign Studies University.

Ludwig M. Eichinger hat das Institut in den Jahren seines Wirkens entscheidend geprägt; in Anerkennung und Dankbarkeit seien ihm diese Bände gewidmet.

Albrecht Plewnia und Andreas Witt
– Reihenherausgeber –

https://doi.org/10.1515/9783110538649-203

Inhalt

Marc Kupietz und Thomas Schmidt

Einleitung

Die Rolle korpusbasierter und korpusgestützer Sprachforschung hat in den letzten 15 Jahren einen Wandel durchlaufen wie wohl kein anderer Bereich der Linguistik. War Korpuslinguistik noch am Anfang der Nullerjahre in Deutschland vornehmlich im Bereich der diachronen Forschung, der Variationslinguistik und der Lexikologie relevant und darüber hinaus teilweise stark von der Computerlinguistik vereinnahmt, so ist sie heute im Mainstream fast aller linguistischer Subdisziplinen fest etabliert. Erfreulicherweise hat sich dabei auch gezeigt, dass qualitative und quantitative linguistische Ansätze sich keineswegs widersprechen, sondern dass gerade ihre Kombination den größten Erkenntnisgewinn verspricht.

Dass das IDS diesen Wandel nicht nur (auch) erfährt, sondern aktiv mitgestaltet hat, zeigt sich unter anderem in der Einrichtung der Programmbereiche „Korpuslinguistik“ (2004) und „Mündliche Korpora“ (2016) und mehrerer Großprojekte in den Abteilungen, etwa „Korpusgrammatik“ (Abteilung Grammatik), „Empirische Methoden“ (Abteilung Lexik) oder „Variation des gesprochenen Deutsch“ (Abteilung Pragmatik), für die korpuslinguistische Zugänge zu Sprache jeweils zentral sind. Nicht zuletzt hat auch der Ausbau der schriftlichen Korpora (Deutsches Referenzkorpus DeReKo) und der mündlichen Korpora (Forschungs- und Lehrkorpus Gesprochenes Deutsch FOLK, Korpora im Archiv für Gesprochenes Deutsch) als Daueraufgabe des Instituts an Bedeutung gewonnen, was sich auch in der Weiterentwicklung der zugehörigen Korpus-Plattformen (KorAP als Nachfolger von COSMAS I/II und die DGD2 als Nachfolger von DIDA und der DGD des DSAv) niederschlägt. Die jeweils 1.000 bis 2.000 neuen Nutzer, die sich jährlich für diese Plattformen registrieren, belegen deutlich die große Verbreitung korpuslinguistischer Methoden in der sprachwissenschaftlichen Forschung und Lehre über das IDS hinaus.

Entsprechend der allgemeinen Etablierung empirischer Verankerung in Korpusdaten und der Verwendung korpuslinguistischer Methodik kann die Aufgabe, einen Überblick über das Themenfeld „Korpuslinguistik“ zu verschaffen, wenn überhaupt nur von der gesamten Reihe geleistet werden. In diesem Band konzentrieren wir uns daher – im Sinne einer Bestandsaufnahme – auf die Grundlagen der Korpuslinguistik (Mair, Kehrein & Vorberger, Dipper & Kwekkeboom, Boas & Fingerhuth, Draxler & Schiel, Schmidt), auf einige interessante neuere Anwendungsfelder (Scharloth, Gries, Imo & Weidner), sowie auf Methodik (Bubenhofer, Scharloth) und Werkzeuge (Dipper & Kwekkeboom, Kehrein & Vorberger, Draxler & Schiel, Schmidt) zur Gewinnung linguistischer Hypothesen und Erschließung linguistischen Wissens aus Korpora.

https://doi.org/10.1515/9783110538649-001

Den Anfang macht Christian **Mair** mit einem Überblick über eine fünfzigjährige *Erfolgsgeschichte Korpuslinguistik?*, der einerseits die zunehmende Verfügbarkeit von Korpusressourcen lobt, andererseits aber auch eine stärkere Einbeziehung gesprochener Spontansprache, den Aufbau mehrsprachiger Korpusressourcen und für die Korpuslinguistik insgesamt eine stärkere Integration mit den Digital Humanities anmahnt und für die Zukunft empfiehlt.

Noah **Bubenhofer** klassifiziert in seinem Beitrag *Visualisierungen in der Korpuslinguistik* diese zunächst nach verschiedenen Grundfiguren und Anwendungszielen, wirbt für das das mittelbare Erkenntnispotential gerade explorativer Visualisierungen als Zwischenschritt des Forschungsprozesses und diskutiert entsprechende Methodologien bzw. Workflows am Beispiel des Einsatzes von Kartendiagrammen zur Exploration von Geokollokationen und anhand der Analyse narrativer Muster in Geburtsberichten mithilfe unterschiedlicher Visualisierungstypen.

Joachim **Scharloths** Beitrag *Korpuslinguistik für sozial- und kulturanalytische Fragestellungen* wirbt für die Übertragung von Prinzipien der *Grounded Theory* auf korpuslinguistische Forschungsprozesse und demonstriert, wie man mit einer entsprechenden datengeleitet korpuslinguistischen Methodik die Analyse von Bundespressekonferenzen zu einem Modell der kommunikativen Gattung führen kann, das zum Verständnis ihrer gesellschaftlichen Funktion und damit zu sozial- und kulturwissenschaftlichem Erkenntnisgewinn beiträgt.

Auch Stefan Th. **Gries** Beitrag *Korpuslinguistik und ihr Potenzial für die (amerikanische) Rechtsprechung* erweitert das Anwendungsfeld korpuslinguistischer Methodik. Er zeigt anhand von zwei bekannten, linguistisch aber zweifelhaften Entscheidungen des Obersten Gerichtshofs der Vereinigten Staaten von Amerika, wie sich mithilfe einer methodisch fundierten Analyse von Konkordanzen und Kollokaten, der im amerikanischen Rechtssystem wichtige und weit verbreitete, aber bisher schlecht definierte Begriff des *ordinary meaning* besser fassen ließe und sich so die amerikanische Rechtsprechung durch die Anwendung korpuslinguistischen Wissens entscheidend verbessern ließe.

Stefanie **Dipper** und Sarah **Kwekkeboom** beschreiben im ersten Teil ihres Beitrags *Historische Linguistik 2.0* die Herausforderungen, die beim Aufbau des Verbundes Historischer Referenzkorpora des Deutschen zutage getreten sind. Im zweiten Teil demonstrieren und diskutieren sie die vielfältigen Möglichkeiten, das online frei verfügbare und detailliert dokumentierte Referenzkorpus Mittelhochdeutsch (ReM) mithilfe von ANNIS im Rahmen historisch-linguistischer Forschung zu nutzen.

Die Bestandsaufnahme zu verschiedenen Typen mündlicher Korpora eröffnen Roland **Kehrein** und Lars **Vorberger** mit ihrem Beitrag zu *Dialekt- und Variationskorpora*. Deren Tradition lässt sich weit in die vordigitale Zeit zurückführen, wo die Dialektforschung bereits im 19. Jahrhundert als eine der ersten

sprachwissenschaftlichen Disziplinen systematische Datenerhebungen zur Erstellung von Dialektatlanten durchgeführt hat. Mit den von Eberhard Zwirner in den 1950er Jahren auf Tonband erhobenen Daten zu den deutschen Mundarten entstanden dann die ersten mit dem Begriff „Korpus" bezeichneten Sammlungen, die für die folgenden Jahrzehnte vorbildhaft waren und ihre moderne Entsprechung in Korpora finden, die auf regionalsprachliche statt „klassisch" dialektale Variation fokussieren. Der Beitrag schließt mit einem Überblick über digitale Instrumente zum Zugriff auf diesen Typus von Korpusdaten.

Hans C. **Boas** und Matthias **Fingerhuth** widmen sich in ihrem Beitrag zu *Deutschen Sprachinselkorpora im 21. Jahrhundert* dann der aktuellen Situation der Dokumentation deutsch(-basiert)er Varietäten außerhalb des deutschsprachigen Kerngebiets. Nach einem historischen Abriss des Forschungsgebiets geben die Autoren einen Überblick über aktuell an verschiedenen Standorten archivierte Sprachinselkorpora und argumentieren für eine stärkere Integration und Koordination dieser verteilten Ressourcen, nicht zuletzt um perspektivisch deren korpuslinguistisches Potenzial besser nutzbar zu machen.

Der Beitrag von Christoph **Draxler** und Florian **Schiel** befasst sich mit der Erstellung und Datenaufbereitung von bzw. für *Moderne Phonetische Datenbanken*. Wie die Autoren zeigen, stehen solche Korpora typischerweise (und in Analogie zur Wechselwirkung zwischen schriftsprachlichen Korpora und Computerlinguistik) an einer Schnittstelle zwischen phonetischer Analyse und angewandter Sprachtechnologie. Es werden zunächst die betreffenden Datentypen beschrieben, bevor anhand eines Beispiels Prozesse der Erstellung und Nutzung eines phonetischen Korpus illustriert werden.

Die Bestandsaufnahme zu mündlichen Korpora beschließt Thomas **Schmidt** mit einem Beitrag zu *Gesprächskorpora*. Nach einer Abgrenzung dieses Korpustyps gegenüber anderen (mündlichen) Korpora werden die wichtigsten seit den 1970er-Jahren erhobenen Gesprächskorpora vorgestellt. Das aktuell im Aufbau befindliche Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) bildet dann den Ausgangspunkt zu einer Diskussion aktueller und künftiger Herausforderungen, die methodologische Innovationen an der Schnittstelle zwischen Korpuslinguistik und Gesprächsforschung notwendig erscheinen lassen.

Im abschließenden Beitrag des Bandes nehmen Wolfgang **Imo** und Beate **Weidner** mit ihrer Diskussion von Einsatzmöglichkeiten *Mündlicher Korpora im DaF- und DaZ-Unterricht* dann eine Anwendungsperspektive ein. Dabei gehen sie zunächst auf die Rolle „authentischer" gesprochener Sprache im Fremdsprachenunterricht ein und stellen mündliche Korpora vor, die für Unterrichtszwecke zur Verfügung stehen. Anschließend wird mit der Plattform Gesprochenes Deutsch eine Datensammlung vorgestellt, die speziell an den Bedürfnissen von Lehrenden und Lernenden in den Bereichen DaF und DaZ ausgerichtet ist.

Christian Mair

1 Erfolgsgeschichte Korpuslinguistik?

Überlegungen zum Fortschritt in der Sprachwissenschaft

Abstract: Der Beitrag gibt einen Überblick über die fünfzigjährige Erfolgsgeschichte der Korpuslinguistik. Er würdigt die enormen technischen Fortschritte und die quantitativ und qualitativ hervorragenden Korpus-Ressourcen, wie sie heute – zumindest für einige große Sprachen – zur Verfügung stehen. Eine direkte Folge der technischen Entwicklung ist der Aufstieg gebrauchsbasierter theoretischer Modelle in der wissenschaftlichen Sprachbeschreibung. Gleichzeitig weist der Beitrag auf die nach wie vor defizitäre Behandlung gesprochener Spontansprache in der Korpuslinguistik hin und empfiehlt, in Zukunft dem Aufbau mehrsprachiger Korpus-Ressourcen mehr Aufmerksamkeit zu schenken als bisher. Abschließend stellt er die Frage nach der zukünftigen Funktion der Korpuslinguistik im Rahmen der *Digital Humanities*.

Keywords: Digital Humanities, gesprochene Sprache, Korpuslinguistik, Mehrsprachigkeit

1 Einleitung

Korpuslinguistik und Konkordanzen haben eine lange Tradition, die weit in die Zeit vor der Digitalisierung zurückreicht. Ohne technische Unterstützung war diese Art von textwissenschaftlicher Analyse allerdings derart aufwendig, dass man die Mühen nur selten auf sich nahm. Konkordanzen wurden typischerweise dann erstellt, wenn – wie im Falle der Bibel oder der Werke Shakespeares – der ideelle Wert des Korpus den Aufwand rechtfertigte.[1] Forscher wie der amerikanische Strukturalist Charles Carpenter Fries, der sich mit derselben Sorgfalt und Andacht einem Korpus von transkribierten Gesprächen von Durchschnittsamerikanerinnen und -amerikanern aus dem Mittleren Westen

1 Als Pionier auf dem Gebiet der Bibelkonkordanz gilt der Dominikaner Hugo von St. Cher (gest. 1263) mit seinen „Concordantiae Sancti Jacobi". Die erste vollständige Konkordanz zu Shakespeares Werken ist Bartlett (1913).

Christian Mair, Englisches Seminar, Universität Freiburg, 79085 Freiburg,
E-Mail: christian.mair@anglistik.uni-freiburg.de

https://doi.org/10.1515/9783110538649-002

widmete, um seine Grammatik des Englischen (Fries 1952) empirisch zu fundieren, blieben die Ausnahme. Fries' Korpus umfasste ca. 250.000 Wörter. Etwas später fasste Randolph Quirk den für damalige Verhältnisse äußerst mutigen Plan, ein eine Million Wörter umfassendes Korpus des britischen Englisch zu erstellen, wobei er außer Tonbandgerät und Schreibmaschine anfangs an keine weiteren technischen Hilfsmittel dachte. Das *Survey of English-Usage*-Korpus wurde allerdings bereits während seiner Fertigstellung in den 1970er Jahren teilweise digitalisiert (Svartvik & Quirk 1980) und ist damit gleichzeitig der letzte Vertreter des korpuslinguistischen Mittelalters und eine Pionierleistung der digitalen korpuslinguistischen Neuzeit.

Bekanntermaßen hat sich die Korpuslinguistik seit diesen Tagen nicht nur in der Anglistik, sondern auch in den anderen Philologien einen zentralen Platz erobert. In Abschnitt 2 werde ich die wesentlichen Etappen einer Entwicklung rekapitulieren, durch die Ideen und Werkzeuge, die von einem kleinen Grüppchen zum Teil beargwöhnter, zum Teil belächelter Exzentriker am Rande ihrer jeweiligen Fächer entwickelt wurden, innerhalb von nur drei Jahrzehnten im Zentrum der Disziplinen angekommen sind. Abschnitt 3 wird sich einigen Schattenseiten dieser Erfolgsgeschichte zuwenden, während der abschließende Abschnitt 4 die Rolle der Technik und des menschlichen Faktors bei der Gestaltung des wissenschaftlichen Fortschritts in der Linguistik abwägt. Der Beitrag schließt mit der Einschätzung, dass sich die Korpuslinguistik gerade wegen ihres Erfolges in den nächsten Jahren im Rahmen der allgemeinen Entwicklung im Bereich der Digitalisierung in den Geistes- und Sozialwissenschaften (Stichwort *Digital Humanities*) neu positionieren muss.

2 *Oh Pioneers!* – Vom Rand in den Mainstream

Als Geburtsjahr der modernen anglistischen Korpuslinguistik gilt gemeinhin das Jahr 1964, in dem W. Nelson Francis, Henry Kučera und ihr kleines Team das Brown Corpus – offiziell „A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers" – fertigstellten. Das Korpus eignete sich in erster Linie als Datengrundlage für gebrauchsbasierte Ansätze zur Beschreibung der amerikanischen Standardvarietät des Englischen, weshalb es in einer Phase, in der die frühe Generative Linguistik Chomskys und die Soziolinguistik Labovs die Sprachwissenschaft in den USA dominierten, nicht gerade auf enthusiastische Rezeption stieß. Für Generativisten waren Korpusdaten überflüssig, weil sie die wahren Herausforderungen für die wissenschaftliche Sprachbeschreibung jenseits der Sammlung relativ willkürlich ausgewählter und strukturell in den meisten Fällen denkbar unauffälliger Performanzdaten

From left: Jostein Hauge, Director of the Norwegian Computing Centre for the Humanities, W. Nelson Francis, Geoffrey Leech, Stig Johansson, Arne Zettersten, Henry Kučera, Randolph Quirk, Jan Svartvik.

Abb. 1.1: Bei der ersten ICAME-Tagung in Bergen (1979), Quelle: http://clu.uni.no/icame/history/First_ICAME_conference_Bergen_1979.gif (letzter Zugriff: 26. 10. 2017).

vermuteten. Für Soziolinguisten wiederum war der schriftsprachliche Standard die denkbar uninteressanteste Varietät, weil hier neben etwas stilistischer Variation nichts zu holen war.

Weitere Initiativen, etwa die offensichtlich sinnvolle Erweiterung der Datengrundlage um ein britisches Parallelkorpus (das LOB [= Lancaster/Oslo-Bergen]-Korpus), kamen nur schleppend voran, so dass die Gründung von ICAME (International Computer Archive of Modern English) im Jahre 1977 in Oslo eher einen ungedeckten Scheck auf die Zukunft denn die Koordination intensiver laufender Forschungsaktivitäten darstellte. Ein Foto von der ersten ICAME Tagung 1979 in Bergen zeigt eine noch eher familiäre Atmosphäre (siehe Abb. 1.1).

Wäre auf dem Bild auch noch John Sinclair zu sehen, hätte man die Pioniere der damaligen anglistischen Korpuslinguistik fast komplett versammelt. Gewisse Hinweise auf den Status und die Reputation des frühen John Sinclair lassen sich allerdings aus der zeitgenössischen Belletristik entnehmen. Dem Vernehmen nach war er die reale Inspiration für den Sprachtechnologen Robin Dempsey aus David Lodges Roman *Small World*, der seine Forschungsprogram-

matik einem staunenden Besucher mit einer Mischung aus Frankensteinschem Fanatismus und missionarischem Eifer vorträgt:

> "I'd like to take you over to our Computer Centre this afternoon," he said. "We've got something set up for you that I think you'll find interesting." He was sort of twitching in his seat with excitement as he said it, like a kid who can't wait to unwrap his Christmas presents. [...] "Anyway," he went on, "when we heard that the University was going to give you an honorary degree, we decided to make yours the first complete corpus in our tape archive." "What does that mean?" I said. "It means," he said, holding up a flat metal canister rather like the sort you keep film spools in, "It means that every word you've ever published is in here." His eyes gleamed with a kind of manic glee, like he was Frankenstein, or some kind of wizard [...] "What's the use of this?" I asked. "What's the use of it?" he said, laughing hysterically, "What's the *use*? Let's show him, Josh." And he passed the canister to the other guy, who takes out a spool of tape and fits it to one of the machines. [...] "What's your favourite word?" "My favourite word? I don't have one." "Oh yes, you do!" he said. "The word you use most frequently." "That's probably *the* or *a* or *and*," I said. He shook his head impatiently. "We instruct the computer to ignore what we call grammatical words – articles, prepositions, pronouns, modal verbs, which have a high frequency rating in all discourse. Then we get to the real nitty-gritty, what we call the lexical words, the words that carry a distinctive semantic content. Words like *love* or *dark* or *heart* or *God*. Let's see." (David Lodge, *Small world*, London: Vintage Books, 2011 [1984], p. 183 f.)

Heutige studentische Leser erkennen die Grundkonstellation: Wortsuche in digitalisierten Texten. Die Details jedoch verwundern. In ein Rechenzentrum gehen zu müssen, um eine triviale Wortsuche in einem doch recht kleinen Korpus, dem publizierten Gesamtwerk eines einzigen Autors, durchzuführen? Datenspeicherung auf filmrollenähnlichem Träger? Wozu der Techniker?

Dieser kurze Abschnitt aus einem Roman macht die beeindruckenden technischen Fortschritte der Korpuslinguistik deutlich. Viele der diesbezüglichen Probleme, mit denen sich Dempsey und sein Techniker noch abmühen mussten, sind heute gelöst. Zumindest für schriftsprachliche Daten ist die Speicherkapazität der Rechner seit Jahrzehnten kein Problem mehr – obwohl die Daten im Gegensatz zum reinen Text der frühen Korpora heute im Regelfall in XML-kompatiblem Format und mit Wortartenannotation vorliegen. Der Zugang zu den meisten Korpora ist niederschwellig möglich – vom Schreibtisch bzw. vom mobilen Endgerät des Nutzers. Die Digitalisierung von Texten ist in der Romanszene noch eine Pioniertat. Was die digitale Erfassung der Textproduktion betrifft, wäre hingegen heute eher die Frage zu stellen, wessen publiziertes Gesamtwerk noch *nicht* digital abrufbar ist.

Auch was Suchabfragen und statistische Analyse der Daten betrifft, sind die Fortschritte enorm. Suchabfrage in digitalen Textdaten ist heute kein Spezialwissen einer kleinen Avantgarde von Experten mehr, sondern essenzielle Kulturtechnik, ohne die niemand bestehen kann, der sich als Teil der moder-

Abb. 1.2: Ngram Viewer-Diagramm „*mountain*, *prairie*", Quelle: http://mentalfloss.com/article/60033/experiments-ngram-art (letzter Zugriff: 17. 10. 2017).

nen Welt versteht. In diesem Sinn finden sich Dempseys Nachfolger heute nicht nur in der Gemeinde der Korpuslinguistinnen und Korpuslinguisten, sondern weit darüber hinaus. Dazu gehört, wer zum Beispiel in harmloser Absicht eine Suchmaske für Informationsabfragen in Datenbanken nutzt, aber auch der wachsende Kreis derjenigen, die im Auftrag diverser Geheimdienste bei der flächendeckenden Überwachung der globalen Kommunikation tätig sind. Letzterer Bereich ist wohl derjenige, in dem in Folge des enormen Datenvolumens tatsächlich noch nicht dezentral auf dem PC gearbeitet werden kann, sondern nach wie vor riesige Rechenzentren notwendig sind. Wenn man will, einer der wenigen Fälle, wo das Gerede von „Big Data in den Geisteswissenschaften" heute nach wie vor zutrifft.[2]

W. Nelson Francis, Geoffrey Leech, Randolph Quirk und John Sinclair hätten vermutlich auch nicht damit gerechnet, dass – fern der geheimnisumwitterten Welt der Nachrichtendienste – Korpuslinguistik zur kreativen Freizeitbeschäftigung wird – zu erkunden auf der Webseite „Experiments in Ngram Art" von Arika Okrent (http://mentalfloss.com/article/60033/experiments-ngram-art [letzter Zugriff: 17. 10. 2017]), die den Google Ngram Viewer nützt, um Diagramme zu erstellen, die auf ikonische Weise die Bedeutung der im Google Books-Material gesuchten Wörter repräsentieren, so etwa im Falle des *mountain*, der sich über der *prairie* erhebt (siehe Abb. 1.2).

2 In diesem Zusammenhang erinnere ich an einen Artikel in der *New York Review of Books* aus dem Jahre 2014, in dem Michael Hayden, der ehemalige Chef der amerikanischen National Security Agency (NSA), mit der denkwürdigen Einsicht zitiert wird, dass „we kill people based on metadata" (Cole 2014).

Doch zurück in den Bereich der Sprachwissenschaft: Korpuslinguistische Verfahren haben innerhalb kürzester Zeit die Lexikologie wie auch die praktische Lexikographie revolutioniert. In der Grammatikforschung waren sie eine wesentliche Voraussetzung für die aktuelle Konjunktur gebrauchsbasierter (*usage-based*) theoretischer Modelle. Weitere wichtige Impulse gab es für die Sprachgeschichte ebenso wie für Textlinguistik und Diskursanalyse. Natürlich erliegt die eine oder andere Studie der Versuchung zum statistischen Positivismus: durchzählen, was sich zählen lässt, und dann sehen wir weiter ...

Intelligente Nutzung der Technik liegt dagegen dann vor, wenn Fragen gestellt werden, die man früher vermieden hätte, weil man keine Hoffnung auf Antwort gehabt hätte. So bedauert der große niederländische Sprachhistoriker Visser noch 1973:

> Today *begin* + form in *-ing* is used with striking frequency alongside of *begin* + infinitive. Which of the two alternatives predominates cannot be ascertained because of the lack of statistical data. (Visser 1970–73, III: 1888)

In der Tat ist die manuelle Auszählung von Gerundien und Infinitiven nach dem Verb *begin* in 20 Millionen Wörtern Text (um eine Zahl zu nennen, die wahrscheinlich der Jahresleistung eines geübten Lesers nahekommt) eine Aufgabe, die man keinem Mitmenschen, auch nicht einem Doktoranden, zumuten möchte. In lemmatisierten und mit Wortartenannotation versehenen Großkorpora wie dem *Corpus of Contemporary American English* (COCA) (Davies 2008–) und dem *Corpus of Historical American English* (COHA) (Davies 2010), mit zusammen fast einer Milliarde Wörtern, kann man innerhalb einiger Stunden ein Vielfaches dieser Textmenge verarbeiten und damit ein recht verlässliches Profil des Sprachgebrauchs erstellen, differenziert nach syntaktischen Umgebungen und Textsorten, und durch Heranziehung weiterer Korpora, natürlich auch nach Varietäten.

Der durchschlagende Erfolg korpuslinguistischer Methodik zeigt sich besonders deutlich, wenn sie von anderen Disziplinen zur Verfolgung eigener Ziele adaptiert wird. So haben Franco Moretti und Mark Algee-Hewitt das Stanford Literary Lab (https://litlab.stanford.edu/[letzter Zugriff: 17. 10. 2017]) aufgebaut, in dem zahlreiche kulturhistorische und literaturwissenschaftliche Fragestellungen nach den Regeln der aktuellen Korpuslinguistik und Texttechnologie erforscht werden. Beispiele für ähnliche Initiativen finden sich natürlich auch in anderen textbezogenen Wissenschaftsdisziplinen. Drei britische Historiker, Clive Emsley, Tim Hitchcock und Robert Shoemaker, haben die Archive des Londoner Zentralgerichtshofs für Strafsachen („Old Bailey“) digitalisiert und stellen ihr Projekt wie folgt vor:

> The Proceedings of the Old Bailey, 1674–1913
> A fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court. (https://www.oldbaileyonline.org/index.jsp [Verson 7.2, letzter Zugriff: 17.10. 2017])

Die Suchmaske dieses Korpus kommt sprachhistorisch interessierten Nutzerinnen und Nutzern nicht entgegen. Sie hilft eher bei der Frage, für welche Art von Verbrechen man zwischen 1788 und 1820 besonders leicht mit Verbannung nach Australien bestraft wurde oder ob sich zwischen Anfang und Ende des 18. Jahrhunderts die Bestrafung von diebischen Hausangestellten weiblichen Geschlechts änderte. Dieses sprachhistorische Defizit ist mittlerweile durch das – von der DFG und später auch im Rahmen von CLARIN-D geförderte – Projekt *Old Bailey Corpus* von Magnus Huber in Gießen umfassend behoben (http://www1.uni-giessen.de/oldbaileycorpus/ [letzter Zugriff: 17.10. 2017] und http://fedora.clarin-d.uni-saarland.de/oldbailey/ [letzter Zugriff: 17.10. 2017]). Korpora fördern eine kooperative Forschungskultur, weil einzelne Nutzer auf den Ergebnissen aller anderen aufbauen können; diese Arbeitsteilung ist effizient und ebnet den Weg zu kumulativem Wissensgewinn. Das Beispiel der doppelten Aufbereitung des Old Bailey-Materials zeigt allerdings, dass diese Art von Teamgeist derzeit noch allzu oft an den Grenzen des jeweiligen Faches endet. Es wäre zu wünschen, dass unter dem interdisziplinären Dach der *Digital Humanities* die digitale Aufbereitung von Sprachdaten in Zukunft auf eine Weise erfolgt, die Anschlussmöglichkeiten in alle relevanten fachlichen Richtungen offen hält und somit ähnliche Duplizierung von Arbeit bereits im Ansatz vermeidet.

3 Die Schattenseiten des Erfolgs

Die oben beschriebene Erfolgsgeschichte hat natürlich auch ihre Schattenseiten. So gebe ich gerne zu, dass mich bislang noch jedes System der Wortartenindizierung in Korpora geärgert hat. Ich weiß, dass häufige Funktionswörter wie *the*, *of* oder *was* in englischen Korpora mit annähernd hundertprozentiger Treffsicherheit korrekt identifiziert werden. Ich zweifle auch nicht an Angaben zur durchschnittlichen Korrektheit des *part-of-speech (POS) tagging*, die sich für standardsprachliches Material im Englischen bei 98 Prozent bewegen. Ebenso weiß ich jedoch aus eigener Erfahrung, dass die entsprechenden Raten bei jedem Thema, das mich wirklich interessiert hat, deutlich unter diesem Wert lagen. Wer zur sprachwissenschaftlich durchaus reizvollen Frage arbeiten will, welche englischen *participles* auch adjektivisch gebraucht werden kön-

nen, sollte sich niemals nur auf _V* und _J*[3] verlassen, sondern zur Sicherheit auf alle Fälle auch die nicht annotierte Version des betreffenden Korpus untersuchen. Zur Probe aufs Exempel und zur Illustration der typischen Probleme möge die folgende kleine Suche nach „STOP annoying_v?g*" im *Corpus of Contemporary American English* (COCA) dienen (also nach Formen des Lemmas *stop*, denen ein als verbales Partizip annotiertes *annoying* folgt). Bei allen fünf Belegen, die auf diese Weise gefunden wurden, war eine adjektivische Lesart von *annoying* (in der Notation dieses Korpus *annoying_JJ*) durch das folgende Wort strukturell blockiert. Sequenzen wie *annoying the people* oder *annoying these people* können im Englischen einfach nicht als Nominalphrase fungieren, weil Artikel und Demonstrativpronomina ausnahmslos *vor* dem attributiven Adjektiv *annoying* erscheinen müssen. Diese Regularität führt zu einem hohen Grad von *Präzision* (*precision*) der Korpussuche: alle Beispiele sind im Sinne der untersuchten Fragestellung brauchbar, und keines muss als irrelevant eliminiert werden. Weniger erfolgreich ist die Suche im Hinblick auf die *vollständige Erfassung* (*recall*) der relevanten Beispiele aus dem Korpus. Wiederholt man dieselbe Suche im nicht annotierten Material („STOP annoying"), findet man ein weiteres Beispiel, und zwar eines, bei dem die verbale Lesart von *annoying* nicht strukturell erzwungen ist, sondern „nur" durch ihre semantisch-pragmatische Plausibilität nahegelegt wird: „That way, they'll soon stop annoying people who are in the process of getting themselves well." Der Kontext erfordert hier die verbale Lesart („Leute zu verärgern") und schließt die adjektivische („lästige Leute") aus; das Annotationsprogramm hingegen vergibt den Adjektiv-Tag (*annoying_JJ*), vermutlich in Analogie zu häufigen Strukturen des Typs *stop stupid people* oder *stop silly nonsense*.

Das gravierendste Defizit in der bisherigen Geschichte der Korpuslinguistik ist jedoch der Umgang mit gesprochenen Daten, wo – wie die folgenden Beispiele zeigen werden – für die nächsten 50 Jahre noch Einiges zu tun bleibt. Abbildung 1.3 zeigt einen der legendären *slips*, mit denen jedes Wort des *Survey of English Usage*-Korpus in monumentalen Zettelkästen mehrfach nach den jeweils relevanten Kategorien abgelegt wurde.

Ohne ins Detail zu gehen, ist deutlich zu erkennen, wie reich diese orthographische Transkription mit prosodischen und paralinguistischen Informationen annotiert ist – und somit innerhalb der Grenzen des Mediums Schrift der Dynamik des gesprochenen Wortes doch einigermaßen gerecht zu werden vermag. Abbildung 1.4 zeigt dieselbe Passage, diesmal in der digitalen Version des gesprochenen Survey-Korpus, die als *London-Lund-Corpus* bekannt ist.

---

3 Hier in der Notation des in Lancaster entwickelten CLAWS-Taggers (http://ucrel.lancs.ac.uk/claws/ [letzter Zugriff: 17.10. 2017]).

```
              S.1.3-9
             *bro*chùre* for# so I q /díd it#q# . and /then ¯ anóther
             *one#p#  - and
          B   *  mhm *
         (A) /Nthèn they *said# well "/now that you've done Nthése#
              and they've been a "/sò sucNcéssful#a# we'd /lı̄ke you
              to do our l Nsùpēr# . /alpha:màtic#l# a or /Nsómething#
m:narrow      and /this is one of Nthése#a# that /goes mm' sídeways#
m':rhythmic   and /fróntwards# and em/bróiders# and */dárns#m'# and
              sews */bùttons on#m#
          B   *( - laughs ) yes*¯
         (A) -- a and I /sāid# well a# I /don't Nrèally *think# I
m:slurred     could /wrı̄te# -- a and this was a m sort of m#a#
              /ninety six page :bòoklet# p /you Nknów# p# about /that
              Nbı̀g# *-* əm I'd I'd /used to gò through# /each of the
          B   * m *
         (A)*processes at :hòme# *.* a I ¯don't think it will be
             *e/nough a# just to have
```

Abb. 1.3: Ausschnitt aus Text S.1.3 (spontanes Gespräch, *Survey of English Usage*-Korpus, Svartvik & Quirk 1980: 16).

BRO☆CHÙRE☆ for■ 139 so I ‖DÍD it■ · 140 and ‖then ANÓTHER one■ –
141 and

b 142 ☆[mhm]☆

>A 141 ‖THÈN they ▷said■ 143 well ˈ‖now that you've done THÉSE■ 144 and they've been ˈ‖SÒ SUCCÉSSFUL■ 145 we'd ‖like you to do our SÙPER■ · 146 ‖ALPHA▵MÀTIC■ 147 or ‖SÓMETHING■ 148 and ‖this is one of THÉSE■ 149 that ‖goes SÍDEWAYS■ 150 and ‖FRÓNTWARDS■ 151 and EM‖BRÓIDERS■ 152 and ☆‖DÁRNS■ 153 and sews☆ ‖BÙTTONS on■

b 154 ☆(– laughs) yes☆

>A 155 – – and I ‖SÃID■ 156 well I ‖don't RÈALLY ▷think■ 157 I could ‖WRĪTE■ – – 158 and this was a sort of ‖ninety-six page ▵BÒOKLET■ 159 ‖you KNÓW■ 160 about ‖that BĪG■ ☆–☆ 161 [əm] I'd I'd ‖need to GÒ through■ 162 ‖each of the

b 163 ☆[m]☆

>A 162 processes at ▵HÒME■ ☆·☆ 164 I don't think it will be e‖nough just to have

Abb. 1.4: Ausschnitt aus Text S.1.3 (spontanes Gespräch, *London-Lund-Corpus* (LLC), Svartvik & Quirk 1980: 85).

Man beachte, dass die 1980 als Buch publizierte digitale Version des Korpus nicht ein direkter Ausdruck des Computer-Texts ist, sondern einige redaktionelle Konzessionen in Richtung bessere Lesbarkeit macht, zum Beispiel durch

Hochstellung der Intonationspfeile von der Grundlinie. Zusätzlich eingeführt ist eine (sprachwissenschaftlich bedeutungslose) laufende Durchnummerierung der Tongruppen. Eliminiert sind dagegen zahlreiche detailliertere prosodische sowie fast alle paralinguistischen Informationen (Stimmqualität etc.). Neuere Standardkorpora wie das COCA bieten zwar wesentlich mehr transkribiertes gesprochenes Material, doch meist aus Sekundärquellen wie Transkripten, die von Rundfunk- und Fernsehanstalten mit Hilfe von linguistisch meist nicht ausgebildetem Personal erstellt werden. Hier als Beispiel ein Interview des Senders ABC mit dem bekannten Country-Musiker Tim McGraw:

(1) TIM-MCGRAW-1SINGE# We met in 1996. It was my first headlining tour and she was my first opening act for my headline tour. And we met then, got married later that year. And we toured again in 2000 for the first Soul2Soul tour. ROBIN-ROBERTS-1AB# (Voiceover) Mm-hmm. TIM-MCGRAW-1SINGE# And last year, we did Soul2Soul II. And, and we decided to do it this year because the kids are getting older... ROBIN-ROBERTS-1AB# (Voiceover) Yeah. TIM-MCGRAW-1SINGE#... and in the summer times, they're not **gonna wanna** go out on the road anymore, so I'll **have to** go by myself in, in the near future. But we thought this might be the last chance that we got to go out together.
(COCA, 2007, ABC_GMA)

In diesem *state-of-the-art mega-corpus* fehlt jede systematische prosodische und paralinguistische Annotation. Das mag zwar die Suchbarkeit erleichtern, die in der digitalen Originalversion des oben erwähnten *London-Lund-Corpus* sehr schlecht war, weil zahlreiche Wörter durch die Markierungen des Intonationsverlaufs zerhackt wurden und das Wort *brochure* (siehe auch oben, Abb. 1.4) deshalb als *bro*ch\ure** erschien, aber natürlich auch in der Form *bro*ch/ure** und zahlreichen anderen Varianten denkbar war.[4] Andererseits schränkt die radikale Reduktion der orthographischen Transkription in COCA die Tauglichkeit des Materials für Forschungen zur gesprochenen Sprache doch deutlich ein. Im Beispiel aus COCA sind die „semi-modals" *have to*, *(have)*

4 Um lexikalische Suchen zu vereinfachen, erstellte ich in den 1990er Jahren für den Freiburger Hausgebrauch eine normalisierte Zusatzversion ohne prosodische Annotation. Vor einer weiteren Verbreitung dieser schreckte ich aus Angst vor einer möglichen negativen Reaktion von Randolph Quirk zurück, der in manchen korpustheoretischen Grundsatzfragen seine Position durchaus emotional vertreten konnte. Diese normalisierte Version wurde auf Datenträgern (*floppy disks*) und mit Konkordanzsoftware (*TACT*) verwendet, die mittlerweile beide obsolet sind und hat sich somit erledigt.

got to, *going to* und *want to* durch Fettdruck hervorgehoben. Die Wahrscheinlichkeit, dass diese verbalen Ausdrücke als lexikalische Einheiten richtig erkannt und transkribiert wurden, ist hoch. Was die phonetische Realisierung betrifft, bleiben jedoch große Unsicherheiten. Die Transkription von *going to want to* mit Hilfe der beiden konventionalisierten Nicht-Standard-Schreibungen *gonna* und *wanna* legt nahe, dass auch tatsächlich Kontraktion vorlag. Der Umkehrschluss trifft nicht zu. Aus der üblichen Schreibung für *have to* darf nicht notwendigerweise auf Abwesenheit phonetischer Kontraktion geschlossen werden. Die entsprechende Schreibweise – *hafta* – ist, im Gegensatz zu *gonna* und *wanna*, noch nicht sehr fest etabliert. Nicht immer hilft Youtube mit Videoausschnitten aus den Originalsendungen, die einem die Überprüfung des gesprochenen Textes ermöglichen würden.

Vielleicht werden wir in zehn Jahren eine Situation erreichen, in der *jedes* gesprochene Korpus multimodal zur Verfügung steht: als Video- oder Tondatei, mit maschineller Unterstützung mit einer orthographischen Transkription aligniert, die in zusätzlichen Versionen mit weiterer grammatischer und prosodischer Information angereichert ist (bzw. von den Nutzerinnen und Nutzern nach eigenen Vorstellungen angereichert werden kann).

Die eklatante Unterrepräsentation spontansprachlicher Daten in Korpora weist auf eine allgemeine Gefahr bei der Entwicklung wissenschaftlicher Infrastrukturen hin. Was sich ohne großen technischen und mit vertretbarem finanziellen Aufwand machen lässt, wird gemacht: und das war in den vergangenen Jahrzehnten die Erstellung immer größerer und immer besser annotierter standardsprachlicher Schriftkorpora. Was nicht in diese Schablone passte, wurde entweder trotzdem in sie gezwängt: man vergleiche etwa die Gewinnung der gesprochenen Daten in COCA aus dem Recycling bereits verfügbarer schriftlicher Quellen, nämlich den orthographischen Transkriptionen von Medienmaterial. Auf jeden Fall verlief die Entwicklung bei der gesprochenen Sprache und ihrer gegenstandsadäquaten korpuslinguistischen Erschließung schleppender. Solche praktischen Zwänge führen dazu, dass man Begleiterscheinungen der Grammatikalisierung in gesprochener Sprache, wie eben die phonetische Reduktion von Hilfsverben, auch heute noch entweder auf der Grundlage von großen Mengen schlecht aufbereiteter Daten oder auf der Grundlage von deutlich zu geringen Mengen qualitativ hochwertiger Daten untersuchen muss. Ein Großkorpus wie COCA enthält zwar große Mengen – in der aktuellen Fassung fast 110 Millionen Wörter – gesprochener Daten, dies jedoch in einem Zustand, der, wie oben gezeigt wurde, phonetische Reduktion nicht annähernd realistisch abbildet. Diejenigen Korpora gesprochener Sprache, die wie das *Santa Barbara corpus of spoken American English* (SBCSAE) (Du Bois et al. 2000–2005) die Anforderungen an die Qualität und Überprüfbarkeit der Tran-

skriptionen erfüllen, stammen aus den 1990er Jahren und umfassen samt und sonders weniger als eine Million Wörter.[5]

Letztendlich ist die Behandlung der gesprochenen Sprache in der Korpuslinguistik ein Beispielfall für eine allgemeinere Problematik. Die Entwicklung von Forschungsinfrastrukturen erfolgt immer im Spannungsfeld zwischen technischer Innovation und der Weiterentwicklung wissenschaftlicher Fragestellungen innerhalb der fachlichen Diskussion. Das Kräfteverhältnis der Beteiligten gestaltet sich dabei allerdings derart, dass beim Aufbau von Forschungsinfrastrukturen oft die technischen Entwickler sowie ihre privaten und öffentlichen Geldgeber die Standards definieren. In diesem Innovationsprozess haben einzelne Nutzer aus den Sprachwissenschaften immer weniger Möglichkeiten, ihre inhaltlich begründete Forschungsprogrammatik als Orientierung für die technische Entwicklung einzubringen. Im schlimmsten Fall führt dies dazu, dass sie, anstatt auf die Anpassung der Infrastruktur an ihre Bedürfnisse zu drängen, ihre eigene Forschungsprogrammatik an den Möglichkeiten der Infrastruktur ausrichten: *linguistics follows technology*. Beispiele für diese Art der Kapitulation vor der Macht des Faktischen sind vielfältig und zahlreich. Eines möchte ich hier herausgreifen und ausführlicher diskutieren.

Die Korpora der postkolonialen Varietäten des Englischen orientieren sich ebenso eindeutig wie unhinterfragt am Konstrukt der nationalen Standardvarietäten des Englischen als plurizentrischer Weltsprache. Der postkoloniale Nationalstaat fungierte bereits beim ersten derartigen Großprojekt als wichtigstes Ordnungskriterium. Das *International Corpus of English* (ICE) wurde Ende der 1980er Jahre von Sidney Greenbaum als ein Cluster von analog strukturieren nationalen Einzelkorpora konzipiert (Greenbaum & Nelson 1996). Heute liegen elf Teilkorpora vollständig vor: Australien, Kanada, Ostafrika (mit Material aus Kenia, Tansania und Malawi), Großbritannien, Indien, Irland, Jamaika, Neuseeland, Nigeria, Philippinen, Singapur.[6] Nicht viel anders ist das seit 2013 verfügbare *Corpus of Global Web-based English* (GloWbE, Davies & Fuchs 2015) strukturiert; es umfasst nationale Komponenten für Großbritannien, die Vereinigten Staaten, Irland, Kanada, Australien, Neuseeland, Indien, Pakistan, Sri Lanka, Bangladesch, Malaysia, Singapur, Philippinen, Hong Kong, Nigeria, Ghana, Kenia, Tansania, Südafrika und Jamaika.

---

5 Das Santa Barbara-Korpus bietet eine prosodisch und diskursanalytisch angereicherte orthographische Transkription mit Zugriff auf die Originaltondateien. Vgl. http://www.linguistics.ucsb.edu/research/santa-barbara-corpus (letzter Zugriff: 17.10. 2017). Das Material umfasst etwa 250.000 Wörter.

6 Weitere Korpora – unter anderem für Fiji, Ghana, Malaysia, Sri Lanka und Trinidad und Tobago – befinden sich in Vorbereitung. Mit ICE-Scotland wird erstmals eine sprachlich eigenständige Region unterhalb der nationalstaatlichen Ebene erschlossen.

Was das überwiegend muttersprachlich gesprochene Englisch Großbritanniens, Irlands, der USA sowie der ehemaligen britischen Siedlerkolonien (Australien, Neuseeland, Kanada ohne Quebec) betrifft, ist das Konstrukt der postkolonialen nationalen Standardvarietät weitgehend unproblematisch. Unstrittig ist auch, dass sich einigermaßen stabile und endonormative nationale Standards des Englischen auf nichtmuttersprachlicher Basis in Afrika (z. B. Nigeria, Ghana, Kenia) und in einigen ehemaligen britischen Kolonien Süd- und Südostasiens (z. B. Indien, Singapur) herausgebildet haben. Wo das nationale Ordnungsprinzip jedoch offensichtlich scheitert, ist Südafrika, weil in diesem Land das muttersprachlich gesprochene Englisch der Nachkommen der britischen Siedler neben den je spezifischen Zweitsprachenvarietäten der Afrikaans-sprachigen weißen Bevölkerung, der indischen Einwanderer sowie der schwarzen Mehrheitsbevölkerung existiert. Hier wie in anderen ethnisch und sprachlich heterogenen Gemeinschaften kann man daher nur hoffen, dass man aus den Metadaten des Korpus (so überhaupt vorhanden) zumindest einen Teil der faszinierenden soziolinguistischen Vielfalt rekonstruieren kann, die sich innerhalb der Grenzen eines Nationalstaats, und innerhalb eines Korpus, findet.

Die additive Reihung von nationalen Teilkorpora suggeriert auch eine egalitäre Vielstimmigkeit, die weit von der Realität der streng hierarchisch geschichteten Plurizentrik der Weltsprache Englisch entfernt ist. Auf dem indischen Subkontinent dominiert das indische Englisch, das über diasporische Verbindungen und durch wachsende mediale Präsenz weit über die Grenzen des Landes ausstrahlt. Die transnationale Wirkung des *Bangladeshi English* ist dagegen beschränkt. Im pazifischen Raum ist Einfluss des australischen Englisch auf das neuseeländische weitaus wahrscheinlicher als Einfluss in der umgekehrten Richtung. Im weltweiten Rahmen überdacht das amerikanische Englisch, zumindest im Bereich des Vokabulars, der Phraseologie und der Grammatik, mittlerweile alle anderen Varietäten, inklusive des britischen Englisch.

Potenziell noch prägender für die Forschung erwies sich bei der Erstellung dieser Korpora der *World Englishes* eine Entscheidung für die Fiktion der Einsprachigkeit als methodisches Prinzip. ICE-Jamaica ist konzipiert als Korpus des jamaikanischen Englisch, obwohl auf der Insel besonders in der gesprochenen Kommunikation ein fließendes Kontinuum zwischen Englisch und Jamaika-Kreol vorherrscht. ICE-India präsentiert sich als englisches Korpus, obwohl Code-Switching zwischen Englisch und anderen indischen Sprachen wie Hindi, Bengali oder Tamilisch in der gesprochenen Kommunikation weit verbreitet ist. Nicht anders verhielt es sich beim philippinischen Teilkorpus (ICE-PHI). Die Projektleiterin benennt in aller Offenheit die erheblichen Probleme, die bei der Erhebung spontaner englischer Gesprächsdaten zu bewältigen waren:

> This is the largest category in the whole of the corpus and, in the Philippine setting, it was a very difficult text type category to collect. Natural conversations among family members and friends in Manila are typically conducted in Tagalog, and, depending on the educational level attained, with little or much English and/or vernacular code-mixing. For purposes of the corpus, the conversations had to be in English with minimal Tagalog insertions, and therefore the collected conversations are already less natural in that respect. The presence of the tape recorder contributed to even less naturalness. (Bautista 2004: 9)

Eine ähnliche „Vorzensur" des Datenmaterials wurde auch beim gesprochenen Textmaterial aus den Medien angewandt, wie die folgende Anmerkung zur Textkategorie „broadcast discussions" belegt:

> The discussions were lifted from television (one discussion was lifted from the radio) and therefore, in general, unless the panelists were talking at the same time, the quality of the recording and of the subsequent transcript is high. In many panel discussions on local television, the language would generally be a code-mixed one. Therefore an important requirement was to identify programs that used English as the matrix language with only minimal switching to Tagalog. (Bautista 2014: 10)

Zumindest für die *Second-Language Varieties of English* erzwingt die Korpus-Schablone somit eine künstliche Homogenisierung der mehrsprachigen Realität; die Kontaktsprachen des Englischen werden entweder bereits im Vorfeld eliminiert (weil allzu mehrsprachige Texte als ungeeignetes Datenmaterial erst gar nicht in Betracht gezogen werden) oder die mehrsprachigen Anteile fristen als *extra-corpus material* ihr Quarantäne-Dasein zwischen den beiden *tags* <indig> und </indig> (für *indigenous*). Dies führt immer dann zu Verzerrungen in der Beschreibung, wenn es spezifische lokale Manifestationen sprachlicher Informalität und stilistischer Variabilität aus der Betrachtung ausklammert.

Als Nutzer macht man andererseits auch bald die Erfahrung, dass trotz aller Bemühungen, das methodisch begründete Desiderat (oder die wirklichkeitsfremde Fiktion?) der einsprachig-englischen ICE-Korpora auch bei den Zweitsprachvarietäten durchzusetzen, doch deutliche Hinweise auf die mehrsprachige Realität erhalten bleiben – wie am folgenden Beispiel aus dem philippinischen Teilkorpus gezeigt werden kann. Mit praktisch allen anderen Varietäten des Englischen teilt das philippinische Englisch die Variation zwischen neutralem bis förmlichem *because* und der umgangssprachlichen Variante *cause*. Im gesprochenen Material des Korpus (ca. 600.000 Wörter) finden sich 2.336 Vorkommen von *because* und 374 von *cause*:

(2) In fact I still have the dust of the Acropolis in my shoes **because** I came directly from <unclear> word </unclear> <,> but one thing I learned from

our short visit to Greece the cradle of democracy and the home of ancient civilization which still influences us to this day <unclear> word </unclear> higher education we must make sure that education does not go to our heads but to our minds <indig> di ko yata [m]aintindihan yun‘a </indig> (<ICE-PHI:S2B-050#5:1:A>)

(3) I you know what I find amazing in college is that **because** since we came from an <.> ex </.> from an exclusive school in Bacolod I mean I find it **’cause** it’s coed so I mean you know the first day that I entered school or I attended classes it’s like oh so this is this is the coed school that my other friends have been experiencing ever since (<ICE-PHI:S1A-075#21:1:B>)

Man beachte, dass Beispiel (2) einen Wechsel ins Tagalog enthält (deutsch: „ich denke, ich halte es für ein“), der nicht durch einen trivialen externen Auslöser – etwa Referenz auf lokale Flora, Fauna, Nahrungsmittel oder Bräuche – bedingt ist. Der oben formulierte programmatische Anspruch auf ein englisches Korpus kann also nur teilweise eingelöst werden.

Mit der Untersuchung der beiden Varianten *because* und *cause*, und auch durch Berücksichtigung der altertümlichen dritten englischen Option, der Konjunktion *for*, ist der Variationsraum bei dieser Variable für das philippinische Englisch noch nicht abgesteckt. Neben *because* und *cause* sind nämlich noch zwei funktional äquivalente Formen aus dem Tagalog einigermaßen regelmäßig belegt, nämlich *kasi* (mit 159 Vorkommen) und *dahil* (23 Vorkommen). *Dahil* ist die traditionelle Form, während *kasi* möglicherweise aus dem Englischen entlehnt ist. Es fällt auf, dass die 23 Vorkommen von *dahil* auf Kontexte beschränkt sind, in denen Tagalog dominiert. *Kasi* dagegen tritt sowohl in solchen Code-switches auf wie in Beispiel (4) als auch als Einzelwortentlehnung wie in Beispiel (5):

(4) <indig> Siguro </indig> come to think of it <indig> siguro <indig> in this case a hidden camera is uh is is good enough <indig> **kasi** walang ano e walang </indig> uh I don’t think the the this uh <unclear> word </unclear> (<ICE-PHI:S1B-038#157:1:B>)

(5) A: Like what bad things
B: Well you know like uhm you know before <indig> **kasi** </indig> it was him and <@> Bianca </@> <indig> e </indig>
A: Yeah

B: Okay they were an item and then suddenly <@> James </@> came into the picture and started telling bad things
A: About Sam
B: About Sam to the people in <@> Bianca </@> ‘s dormitory
A: Oh really
(<ICE-PHI:S1A-097#45:1:A> bis <ICE-PHI:S1A-097#51:1:A>)

Beispiel (5) zeigt auf jeden Fall, dass es neben der internationalen informellen Variante *cause* im philippinischen Englisch auch noch eine spezifisch lokale Möglichkeit gibt, Informalität zu signalisieren, nämlich *kasi*. Über den Einzelfall hinaus zeigt das Beispiel, dass es verfehlt wäre, in scheinbar unordentlichen Korpusdaten aufräumen zu wollen, indem *kasi* in Beispiel (5) als Lehnwort und somit Teil des philippinischen Englisch klassifiziert, in Beispiel (4) jedoch als einen Anglizismus im Tagalog. Vielmehr geht es darum zu zeigen, dass eine vollständige Beschreibung des philippinischen Englisch nur im mehrsprachigen Kontext möglich ist und eine neue Generation von ICE-Korpora diesem Umstand in ihrem Design vielleicht offensiver Rechnung tragen könnte.

Im ICE-Korpus, wie es konzipiert wurde, ist eine Passage wie in Beispiel (6) ein Betriebsunfall, der auf der Grundlage des theoretisch formulierten Anspruchs nicht hätte passieren sollen:

(6) <indig> Ngayong hapon mga kaibigan nakita niyo po yung </indig> medal tally <indig> mukhang magiging </indig> exciting <indig> yung labanan ng US at China dahil tatlong </indig> gold medals <indig> na lang ang uh lamang ng US sa China </indig> so any time any moment <indig> ay pwedeng uh mapantayan or malampasan ng China ito dahil </indig> ongoing <indig> pa rin po ang mga </indig> gold medal matches <indig> dito po sa </indig> twenty-seventh Olympiad
(<ICE-PHI:S2A-019#3:1:A>)

In einem noch zu erstellenden mehrsprachigen *Corpus of Educated Philippine Spoken Usage* könnten auf der Grundlage solcher Daten hingegen völlig neue Einblicke in eine komplexe sprachliche Realität gewonnen werden.

Die Französische Revolution wird oft dafür verantwortlich gemacht, dass sich in ihrer Folge in Europa und in den von ihm kolonisierten Teilen der Welt eine Ideologie der Einheit von Sprache und Nation verbreitet habe. Angesichts des Überwiegens einsprachiger Korpora und der nach wie vor starken nationalen Traditionen in der Korpusforschung ließe sich durchaus die Frage stellen, ob der lange Schatten der Französischen Revolution bis in die heutigen digitalen Forschungsinfrastrukturen reicht: Eine Nation mit einer Sprache braucht

ein einsprachiges Nationalkorpus.[7] Vielleicht ist die Zeit gekommen, die Dominanz einsprachiger schriftorientierter Standard-Korpora in Frage zu stellen und die Arbeit mit Nichtstandardvarietäten, mit gesprochener Spontansprache und mit mehrsprachigen Daten als neue Forschungsprioritäten der Korpuslinguistik zu etablieren. Die technischen Entwickler und die Geldgeber könnten nach dieser Vorgabe bei der Umsetzung der neuen Forschungsagenda helfen. Dies wäre eine Umkehrung der bisher üblichen Arbeitsabläufe, bei denen sich Wissenschaftlerinnen und Wissenschaftler allzu oft an den technischen Gegebenheiten ausrichten mussten, doch würde es an innovativen, wissenschaftlich anspruchsvollen und auch gesellschaftlich nützlichen Forschungsvorhaben nicht mangeln.

4 Wissenschaftlicher Fortschritt im Spannungsfeld zwischen Mensch und Technik

Wenn wir auf die Geschichte der Linguistik zurückblicken, können wir feststellen, dass manche wissenschaftliche Revolutionen im Fach von der technischen Entwicklung weitgehend entkoppelt verliefen. So benötigte Ferdinand de Saussure für die Konzeption des strukturalistischen Ansatzes wohl nicht mehr als Papier und Bleistift. Soziolinguistik und Diskursanalyse andererseits wären ohne die Erfindung von Tonaufnahmetechniken und der damit ermöglichten Speicherung akustischer Daten wohl nicht denkbar. Und jede Verbesserung der technischen Ausstattung – z. B. Erhöhung der Aufnahme- und Wiedergabequalität oder Verkleinerung der Geräte für den mobilen Einsatz – erweitert sofort den Horizont für die wissenschaftliche Arbeit. Michael Halliday hat insofern nur wenig übertrieben, wenn er feststellt:

> Perhaps the greatest single event in the history of linguistics was the invention of the tape recorder, which for the first time has captured natural conversation and made it accessible to systematic study. (Halliday 1994: xxiii)

Einen ähnlichen Innovationsschub verdankt die Linguistik auch dem Computer. Einerseits funktioniert all das, was man mit Tonband und Videorekorder machen kann, mit digitalen Verfahren noch besser und schneller. Andererseits

---

7 Das *British National Corpus* (BNC) ist in dieser Hinsicht eine Fehlbenennung, weil die englische Sprache auf so offensichtliche Weise größer ist als die Summe der Nationalstaaten, in denen sie als Muttersprache gesprochen wird. Das *Russian National Corpus* ist ein Grenzfall. Das ungarische, polnische und tschechische Nationalkorpus entsprechen dem Programm.

entwickelten sich – zumindest für die kleine Zahl der Sprachen, die traditionell die größte wissenschaftliche Aufmerksamkeit auf sich ziehen – sehr rasch luxuriöse korpuslinguistische Arbeitsumgebungen, die an Fülle und Diversität des Datenmaterials alles Frühere bei Weitem in den Schatten stellten.[8]

Resultat war ein ungeahnter Aufschwung von gebrauchsbasierten Beschreibungsansätzen und statistischen Analyseverfahren. Dort, wo sich technische und konzeptuelle Innovation im Gleichgewicht befinden, herrschen fast paradiesische Zustände. Bei vielen Fragen haben sich die Mühen, die vor der Digitalisierung mit der Datenbeschaffung und -aufbereitung verbunden waren, drastisch verringert. „First throw away your evidence!" war der Slogan, mit dem John Sinclair (1986) die neue linguistische Arbeitspraxis plakativ beschrieb. Korpora ermöglichen neue datengetriebene (*bottom up*) Ansätze und erhöhen das Potenzial und die Effizienz etablierter konzeptuell-theoretisch motivierter (*top down*) Verfahren. Einen guten Beleg dafür liefert die Vielzahl jüngerer korpuslinguistischer Arbeiten zur Grammatikalisierung, in der ein über hundert Jahre alter theoretischer Begriff in einem neuen technischen Umfeld wieder zum Strahlen gebracht wird (vgl. den Forschungsüberblick in Mair 2011).

Ein letzter Aspekt sollte bei der Bilanz nicht vergessen werden. Quer über alle Einzelthemen hinweg führen Korpora auch zu einer neuen Arbeitskultur in der Sprachwissenschaft. Sie fördern kooperative Forschung und kumulativen Erkenntnisgewinn, weil viele Wissenschaftlerinnen und Wissenschaftler von einer gemeinsamen Datengrundlage ausgehen. Was die Zukunft des Fachs betrifft, muss klar sein, dass gerade die mehrsprachigen Korpora, deren Erstellung oben als Desiderat formuliert wurde, nur in Teamarbeit kompiliert und analysiert werden können. Korpora haben auch das Potenzial, gewisse (sprach)wissenschaftliche Hierarchien zu unterminieren, so etwa diejenige zwischen Linguisten als *native speaker* und *non-native speaker* der untersuchten Sprache (wovon ich als deutschsprachiger Anglist selber profitiert habe). Und schließlich gelangt der Nachwuchs in der Korpuslinguistik oft sehr früh zu Autorität: durch Technikaffinität und Vertrautheit mit komplexen Verfahren der statistischen Analyse und Visualisierung.

8 Ich erinnere an die Einschränkung, die ich oben für spontansprachliche Daten gemacht habe.

5 Die Selbstauflösung der Korpuslinguistik und ihre Neukonzeption als Teil der *Digital Humanities*

Nach einer fünfzigjährigen stürmischen Entwicklung mit insgesamt sehr positiver Bilanz zeichnet sich für die Korpuslinguistik der nächste Umbruch ab. Bis in die frühen 1990er Jahre benutzten nur sehr wenige Sprachwissenschaftlerinnen und Sprachwissenschaftler digitale Korpora, und die Mehrheit der *Community* hatte wenig Ahnung, wozu solche Werkzeuge gut sein sollten und wie sie im Detail funktionierten. Diese Situation hat sich grundlegend gewandelt. Korpora sind überall. Der Einstieg in die Korpuslinguistik ist niederschwellig möglich. Der prinzipielle Nutzen der neuen Verfahren wird nicht in Frage gestellt. Gerade wegen des Überangebots und der raschen technischen Entwicklung ist es heute für Einsteiger jedoch schwieriger geworden, ein Verständnis dafür zu entwickeln, wie die Werkzeuge im Detail funktionieren und was ihre Potenziale und Grenzen sind.

In gewisser Weise ist der Korpuslinguistik sogar ein zentrales Identifikationsmerkmal abhandengekommen, nämlich der Konsens darüber, was ein Korpus ist. Die enge Definition des linguistischen Korpus, als digitale Textsammlung, die von Sprachwissenschaftlern für die Zwecke sprachwissenschaftlicher Analyse erstellt wurde, gilt nur mehr bedingt. Viele Korpora werden nicht mehr nach traditionellen Verfahren kompiliert und dann digitalisiert, sondern durch Recycling bestehender Digitalisate erzeugt. Nicht nur die Linguistik, sondern alle Wissenschaften, die mit sprachlichen Daten arbeiten, operieren in einem Universum von digitalisierten Texten, in das zunehmend auch multimodale Sprachdaten aufgenommen werden. Um mit dem Bild zu sprechen, das Gatto (2011) geprägt hat: Die feste Grundlage eines klar definierten Korpus („body") ist verloren gegangen, und an ihre Stelle ist ein vielfältiges, reiches, aber auch verwirrend unübersichtliches digitales Sprachnetz getreten („web").

„Einführungen in die Korpuslinguistik" für Studierende im Grundstudium[9] wird und soll es auch in Zukunft geben. Eines ihrer wichtigen Lehrziele wird jedoch sein, ein Bewusstsein dafür zu schaffen, dass die interessantesten digitalisierten Sprachdaten immer öfter nicht in Form traditioneller Korpora vorliegen werden und dass die Linguistik ihre vormalige Führungsrolle bei der krea-

---

9 Vgl. z. B. Gerstenberg (2013) für die Romanistik, Lemnitzer & Zinsmeister (2010) bzw. Perkuhn, Keibel & Kupietz (2012) für die Germanistik und McEnery & Hardie (2012) für die Anglistik.

tiven wissenschaftlichen Arbeit mit Korpora und digitalen Sprachdaten zunehmend mit anderen Disziplinen teilen wird müssen. Am Ende einer beeindruckenden Erfolgsgeschichte ist die Korpuslinguistik somit dabei, sich aufzulösen und in die *Digital Humanities*-Bewegung zu integrieren.

Literatur

Bartlett, John (1913): *A complete concordance or verbal index to words, phrases, and passages in the dramatic works of Shakespeare*. London: Macmillan.

Bautista, Ma. Lourdes (2004): An overview of the Philippine component of the International Corpus of English. *Asian Englishes* 7, 8–26.

Cole, David (2014): We kill people based on metadata. *New York Review of Books*, 10. Mai 2014. http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark (2008–): *The Corpus of Contemporary American English (COCA): 520 million words, 1990–present*. https://corpus.byu.edu/coca/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark (2010–): The Corpus of Historical American English (COHA): 400 million words, 1810–2009. https://corpus.byu.edu/coha/ (letzter Zugriff: 17. 10. 2017).

Davies, Mark & Robert Fuchs (2015): Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-Based English Corpus (GloWbE). *English World-Wide* 36, 1–28.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson & Nii Martey (2000–2005): *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia, PA: Linguistic Data Consortium.

Fries, Charles C. (1952): *The structure of English: The construction of English sentences*. New York: Harcourt, Brace.

Gatto, Maristella (2011): The ‘body’ and the ‘web’: The web as corpus ten years on. *ICAME Journal* 35, 35–58.

Gerstenberg, Annette (2013): *Arbeitstechniken für Romanisten*. Berlin, Boston: de Gruyter Mouton.

Greenbaum, Sidney & Gerald Nelson (1996): The International Corpus of English (ICE) project. *World Englishes* 15, 3–15.

Halliday, M. A. K. (1994*): An introduction to functional grammar*. 2. Aufl. London: Edward Arnold.

Lemnitzer, Lothar & Heike Zinsmeister (2010): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Mair, Christian (2011): Grammaticalization and corpus linguistics. In Heiko Narrog & Bernd Heine (Hrsg.), *The Oxford handbook of grammaticalization*, 239–250. Oxford: Oxford University Press.

McEnery, Tony & Andrew Hardie (2012): *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012*): Korpuslinguistik*. Paderborn: Fink.

Sinclair, John (1986): First throw away your evidence! In Gerhard Leitner (Hrsg.), *The English reference grammar: Language and linguistics, writers and readers*, 56–64. Tübingen: Niemeyer.

Svartvik, Jan & Randolph Quirk (Hrsg.) (1980): *A corpus of English conversation* (Lund Studies in English 56). Lund: Liber/Gleerups.
Visser, Frederikus Th. (1970–1978): *An historical syntax of the English language*. 3 Bde. Leiden: Brill.

Noah Bubenhofer

2 Visualisierungen in der Korpuslinguistik

Diagrammatische Operationen zur Gegenstandkonstitution, -analyse und Ergebnispräsentation

Abstract: Visualisierungen sind auch in der Korpuslinguistik wichtig, um Strukturen in großen Korpora überhaupt analysierbar zu machen. Daher sind Methoden der „Visual Analytics" nicht einfach der letzte Schritt einer Korpusanalyse, sondern beeinflussen bereits die Datenaufbereitung. Aus Sicht der Diagrammatik, der Lehre des Diagramms, lässt sich gut herleiten, warum Visualisierungen eigentliche „Denkzeuge" sind: Mit Diagrammen kann operiert werden und im besten Fall können aus bestehendem Wissen neue Erkenntnisse gewonnen werden. Für den korpuslinguistischen Zugang sind einige sog. diagrammatische Grundfiguren, also grundlegende Typen von Diagrammen, entscheidende Mittel, um den Untersuchungsgegenstand Sprache zu konstituieren, so z. B. die Liste, der Vektor und der Graph. Der Beitrag konzipiert diagrammatische Operationen als Grundbedingung der Korpuslinguistik und skizziert fünf diagrammatische Grundfiguren. Zusätzlich wird an einem Beispiel, der Analyse von sog. Geokollokationen, gezeigt, worin der Wert explorativer visueller Korpusanalysen besteht.

Keywords: datengeleitete Korpuslinguistik, Datenvisualisierung, Diagrammatik, Visual Analytics

1 Visualisierung als Diagramm

Die klassische Keyword in Context-Liste (KWiC) oder Konkordanz ist in der modernen Korpuslinguistik häufig nicht mal mehr der Startpunkt einer Analyse, sondern wird oft gleich in eine Tabelle oder ein Diagramm übersetzt, das die

Anmerkung: Der Text entstand im Rahmen des vom Schweizer Nationalfonds geförderten Projektes „Visual Linguistics", bei dem auch die folgenden Personen mitarbeiteten: Klaus Rothenhäusler, Irene Ma, Danica Pajovic und Katrin Affolter.

Noah Bubenhofer, Institut für Computerlinguistik, Universität Zürich, Andreasstr. 15, CH-8050 Zürich, Schweiz, E-Mail: bubenhofer@cl.uzh.ch

https://doi.org/10.1515/9783110538649-003

Verteilung eines Phänomens beschreibt. Trotzdem ist die KWiC-Liste nach wie vor ein Ikon für den korpuslinguistischen Zugang – und ich möchte im Folgenden argumentieren, dass sie weit mehr ist als das. Es handelt sich bei der Kompilation einer KWiC-Liste um eine der grundlegendsten diagrammatischen Operationen, die den korpuslinguistischen Zugang zu Sprachdaten von anderen Zugängen der Lektüre oder Analyse unterscheidet. Die KWiC-Liste ist eine Form eines Indexes oder Registers, mit dem die Einheit des Textes aufgebrochen wird. Dies ist einerseits ein Verlust, da die ursprüngliche Komplexität der Textgestalt nicht mehr (oder nur ansatzweise) sichtbar ist. Andererseits ist dies jedoch ein enormer Gewinn, weil damit eine neue Sicht auf Texte und auf die Serialität von bestimmten Phänomenen in den Texten gewonnen wird.

Es ist ein Gemeinplatz in der Korpuslinguistik, dass diese neue Perspektive auf Textkorpora gewinnbringend ist, um die Musterhaftigkeit von sprachlichen Phänomenen zu entdecken (Perkuhn & Belica 2006; Perkuhn, Keibel & Kupietz 2012). Da ich im vorliegenden Beitrag jedoch den Zusammenhang von Visualisierungen und Korpuslinguistik beleuchte, möchte ich mit einer diagrammatischen Sicht auf korpuslinguistische Analyseformen beginnen.

1.1 Die diagrammatische Operation als Grundbedingung der Korpuslinguistik

Die Theoriebildung zur Diagrammatik (Bauer & Ernst 2010; Krämer 2016; Stetter 2005; Bredekamp 2008) geht insbesondere auf Arbeiten von Charles Sanders Peirce zurück, der mit der Trias von Index, Ikon und Symbol das Erscheinungsspektrum von Zeichen beschrieben hat. Das prototypische Diagramm, beispielsweise die geometrische Zeichnung eines Dreiecks oder Datenpunkte in einem Koordinatensystem, steht zwar in einem ikonischen Ähnlichkeitsverhältnis zum Denotat, abstrahiert davon jedoch, so dass es als „Verwirklichung eines abstrakten Modells“ (Eco 1977: 55) wahrgenommen werden kann. Auch wenn das Dreieck nicht genau wie ein Dreieck aussieht (etwa, weil es von Hand skizziert worden ist), wird es in bestimmten Kontexten als abstraktes Modell des Dreiecks wahrgenommen. Wenn es so wahrgenommen wird, wird es als Diagramm – und nicht etwa als Skizze, Gemälde o. ä. wahrgenommen, wo die Strichdicke, Farbe, Strichqualität etc. als bedeutungstragend aufgefasst würde.[1]

1 Natürlich können die genannten Merkmale auch bei einem Diagramm bedeutungstragend sein, etwa um verschiedene Typen von Dreiecken zu unterscheiden. Aber auch dann wird beispielsweise die Farbe Rot nicht als „rot“ wahrgenommen, sondern in ihrer diagrammatischen Funktion im Diagramm.

Mit Peirce gesprochen würden solche Merkmale nicht als „Qualizeichen“, sondern als „Sinzeichen“ aufgefasst (Peirce 1994: 2244).

Das Dreieck als Diagramm ist also *schematisch* und damit „auf Reproduzierbarkeit hin angelegt“ (Krämer 2016: 76). Es wird in Form einer grafischen Umsetzung instantiiert und nur in dieser Instanz sichtbar.

> Alles, was schematisch ist, kann wiederholt und in dieser Wiederholung – absichtsvoll oder versehentlich – zugleich variiert werden. Dies ist ein Grundzug aller diagrammatischer Artefakte. (Krämer 2016: 77)

Zum Schematismus kommen jedoch einige weitere diagrammtypische Eigenschaften hinzu (Sybille Krämer nennt deren zwölf, Krämer 2016: 60 ff.), darunter insbesondere solche, die mit der Ausbreitung von Informationen auf einer Fläche, auf der „operiert“ werden kann, zu tun haben. Diagramme benutzen interagierende *grafische Ausdrucksformen* (Punkt, Linie, Fläche, Text etc.), die auf einer *gerichteten Fläche* „unterschiedliche Gesichtspunkte und Ansichten“ (Krämer 2016: 74) simultan darstellen. Diese Mittel ermöglichen es, „Relationen mit Hilfe von Relationen“ (Krämer 2016: 70) darzustellen, etwa beim Dreieck, indem die Positionen der Katheten zueinander und in Relation zur Hypotenuse sichtbar werden. Denkt man an ein Punktdiagramm, wird dabei noch deutlicher, dass damit eine Verräumlichung einhergeht, in der nichträumliche Daten (Zahlenwerte) auf einer gerichteten Fläche in Relation zueinander grafisch dargestellt werden. Also: „Räumliche Relationen artikulieren – zumeist – nichträumliche Relationen“ (Krämer 2016: 71), wobei Karten, Grundrisszeichnungen etc. Ausnahmen bilden. Gerade in der Korpuslinguistik haben wir es aber bei Diagrammen sehr häufig mit einer verräumlichten Darstellung zu tun – oder zumindest mit einer anders räumlich organisierten Darstellung, indem etwa die Sequenzialität von Texten aufgebrochen und vom syntaktischen in einen anders modellierten Raum überführt wird. Wenn Kollokationen beispielsweise als Netzgraph dargestellt werden, orientiert sich die Ordnung der Kollokatoren im Raum am statistischen Distributionsverhalten der Kollokationen im Korpus.

Die diagrammatisch vielleicht unscheinbare KWiC-Liste zeigt die erwähnten Eigenschaften eines Diagramms jedoch deutlich: Textfundstellen werden als Liste, damit also in Form von Zeilen ausgegeben, die in Relation zueinander, auf einer gerichteten Fläche (von oben nach unten zu lesen) stehen. Die Zeilen bilden damit einen Raum. Durch eine bestimmte Sortierung (nach Fundstellen, alphabetisch nach Kontext o. ä.) kann die Relationierung der Zeilen zueinander verändert werden. Die instantiierte KWiC-Liste verweist auf ein abstraktes, immaterielles Schema „Liste“, mit dem, bei entsprechendem Wissen oder entsprechender Anleitung, die Liste interpretiert wird. Und in ihrer Form, der verräumlichten Darstellung, stellt die Liste eine Synopse dar, mit

der die an völlig unterschiedlichen Stellen auftretenden Belege simultan untereinander erscheinen.[2]

Entscheidend für die linguistische Interpretation der so dargestellten Daten ist dabei, dass mit dieser Liste – mit Sybille Krämer gesprochen – „operiert“ werden kann:

> Gleich einer Karte, welche Bewegungen in einem unvertrauten Terrain eröffnet, ermöglichen Diagramme, dass wir praktisch oder theoretisch etwas tun, was ohne Diagramm schwer oder überhaupt nicht auszuführen ist. Diagramme sind graphische Denkzeuge; sie eröffnen kognitive Bewegungsmöglichkeiten, insofern ihrem Gebrauch ein transfiguratives Potenzial innewohnt, kraft dessen graphische Konstellationen und deren handgreifliche Manipulation als intellektuelle Tätigkeiten interpretierbar werden. (Krämer 2016: 83)

Die Transfiguration der KWiC-Liste, also die räumliche, gerichtete Anordnung von Textbelegen, ermöglicht beispielsweise (bei geeigneter Datenlage) den interpretativen Schritt, daraus den Wortgebrauch und damit die Semantik eines Lexems, eine grammatische Regularität etc. abzuleiten. Das Diagramm alleine ist dafür nicht ausreichend, es benötigt auch einen theoretischen Standpunkt, von dem aus sich diese Interpretation motivieren lässt, es benötigt eine Möglichkeit, die Liste zu erstellen, aber das Schema der Liste als Diagramm ist die *conditio sine qua non*, um überhaupt eine korpuslinguistische Analyse zu ermöglichen.

Um die interpretativen Schritte zu ermöglichen, muss mit der KWiC-Liste „operiert“ werden. Dies ist möglich, indem einerseits dank ihrer Diagrammhaftigkeit, in der die Fundstellen in Relation zueinander stehen und synoptisch auf einer gerichteten Fläche dargestellt werden, diese Relationen gelesen, gedeutet und manipuliert werden können. Es ist möglich, die Sortierung nach zahlreichen Kriterien zu ändern, Gruppen zu bilden und dergleichen. Andererseits stellt das Diagramm etwas dar, weist also eine Referenzialität zu etwas Gemeintem, einem diagrammexternen Sachverhalt her, überführt aber die Relationen dieses Sachverhalts in das System des Diagramms, das im Fall der KWiC-Liste dem Schema der Index-Liste gehorcht. Die Annahme, dass eine Beobachtung im Diagramm eine Parallelität zum Sachverhalt aufweist, macht die Arbeit mit dem Diagramm ja überhaupt erst sinnvoll. Man würde eine Landkarte oder einen Stadtplan nicht benutzen wollen, wenn man nicht der Überzeugung wäre, dass die Karte bzw. der Plan die wirkliche Landschaft oder Stadt

---

2 Die Homophonie von „KWiC“ zum englischen „quick“ verleitete den Namensgeber des Terminus, Luhn (1960) (Manning & Schütze 2002: 35), wahrscheinlich auch zu diesem Akronym und verweist genau auf die Funktion der Synopsis, mit der ein schneller Überblick versprochen wird.

in einer Art und Weise abbildet, so dass man sich dank des Plans darin orientieren kann. Dies macht aber deutlich, dass Diagramme „graphische Denkzeuge“ (Krämer 2016: 83) sind – nicht nur im Fall der Korpuslinguistik an entscheidender Stelle einer ganzen Methodologie.

Die KWiC-Liste ist zwar ein bedeutendes Element korpuslinguistischer Analyse, jedoch nicht das einzige. Im Gebrauch ist eine Vielzahl von weiteren Diagrammtypen, die auf wichtige *diagrammatische Grundfiguren*, wie ich sie nennen möchte, zurückgehen. Visualisierungen in der Korpuslinguistik zu thematisieren, bedeutet vor dem Hintergrund der Diagrammatik also anzuerkennen, dass diese nicht einfach schönes Ornat für wissenschaftliche Publikationen sind, auch mehr als Analysewerkzeuge, wenn man die Perspektive der Visual Analytics hinzuzieht, sondern in ihren diagrammatischen Grundfiguren entscheidende Schritte der Gegenstands- und Methodenkonstitution, ohne die die Korpuslinguistik nicht wäre, was sie ist.

Im Folgenden möchte ich deshalb kurz auf typische Formen der Visualisierung in der Korpuslinguistik eingehen und vorschlagen, auf welche diagrammatischen Grundfiguren sie zurückzuführen sind. Dann ist es möglich herauszuarbeiten, wie diese Grundfiguren die Gegenstände und Analysemethoden konstituieren, was im vorliegenden Beitrag jedoch nur angedeutet werden kann.[3] Dieser Abschnitt gibt mir auch die Gelegenheit, auf einige wichtige Arbeiten am Institut für Deutsche Sprache (IDS) im Bereich der Datenvisualisierung aufmerksam zu machen.

Anschließend werde ich zwei Beispiele explorativer Visualisierungen zeigen. Beenden möchte ich den Beitrag mit einem Plädoyer für mehr diagrammatische Experimentierfreude.

1.2 Figuren der Visualisierung in der Korpuslinguistik

In der Linguistik generell, insbesondere aber auch in der Korpuslinguistik, sind m. E. folgende diagrammatischen Grundfiguren besonders relevant: Die Liste (mit allen Sonderformen wie z. B. der Tabelle), die Karte, die Vektoren, der Graph und die Partitur (vgl. Abb. 2.1). Die Typen Karte, Vektoren (z. B. in Form von Achsendiagrammen) und Graph werden in vielen Diagrammklassifizierungen genannt (so etwa im Wikipediaeintrag zu „Diagramm“), die Liste und die Partitur werden normalerweise jedoch nicht dazu gezählt. Zudem geht es mir weniger um eine formale Unterscheidung der Typen, sondern um die Frage,

3 Ich verweise stattdessen auf Bubenhofer (2018c); Bubenhofer et al. (2017); Bubenhofer & Rothenhäusler (2016).

Abb. 2.1: Fünf für die Linguistik bedeutende diagrammatische Grundfiguren: Liste, Karte, Partitur, Vektoren, Graph.

wie diese Diagrammtypen sozusagen als Denkfigur in der Linguistik gegenstandskonstituierend wirken.

Im Abschnitt 1.1 habe ich die Bedeutung der KWiC-*Liste* in der Korpuslinguistik bereits betont. Generell ist die Liste eine diagrammatische Grundfigur, die für viele Wissensbereiche von großer Bedeutung ist (Echterhölter 2015; Jullien 2004; Pigeot 2004; Eco 2009). Als Index-Liste, die Fundstellen aus verschiedenen Textstellen vereint und auf die entsprechenden Quellen verweist, ist sie die Grundfigur korpuslinguistischen Arbeitens. Der Theologe und Linguist Roberto Busa war einer der Ersten, der die rechnergestützte Indexerstellung mittels IBM Lochkartenrechnern verwendete, um seinen *Index Thomesticus* zu erstellen (Busa 1951; Bonfanti 2012). Die Grundfigur erwies sich aber lange vor den ersten Computern als hilfreich, etwa für die Erstellung von Enzyklopädien (Placcius 1689; Siegel 2009).

Listen können sehr unterschiedliche Formen annehmen. Ein interessantes Beispiel ist das Kollokationsprofil, eine Liste von Lexemen, die statistisch signifikante Kollokatoren zu einem Basislexem sind. Besonders erwähnenswert ist die von Cyril Belica entwickelte Kookkurrenzberechnung (vgl. Abb. 2.2), die zwar nach dem Vorbild der Kollokationsanalyse vorgeht, jedoch weitere vielfältige Informationen in der Liste der Kollokatoren vereint: weitere sekundäre und tertiäre Kollokatoren, typische syntagmatische Muster mit Angaben zu deren Verbreitung etc. (Belica 2001 ff.).

Karten genießen in wissenschaftlichen Visualisierungen generell einen hohen Stellenwert und in der Linguistik sind sie z. B. in der Dialektologie bereits lange gebräuchlich, wo sie sowohl „Dokumentations-“ als auch „Forschungsmittel“ (Naumann 1982) sind. Sie dienen also sowohl der Präsentation als auch der Exploration der Daten (siehe dazu Abschnitt 1.3). Mit einer Korpusgrundlage lassen sich Karten, sofern Georeferenzen vorhanden sind, problemlos automatisch erstellen (siehe dazu auch Abschnitt 2.1). Damit werden sie für diatopische Fragestellungen zu einem wichtigen Mittel der Datenexploration, wie z. B. das Projekt „Gesprochenes Deutsch“ am Institut für

Analysewort: Ausländer, Analysetyp 0

-1	-1	19509	**lebenden** hier legal	9	44%	von legal hier lebenden Ausländern
-1	-1	19509	lebenden hier rechtmäßig	11	54%	sollten alle rechtmäßig hier lebenden Ausländer auch arbeiten
-1	-1	19509	lebenden hier	545	61%	der|die hier [...] lebenden [...] Ausländer
-1	-1	19509	lebenden legal	51	52%	von legal [in Österreich] lebenden Ausländern
-1	-1	19509	lebenden rechtmäßig	33	51%	von rechtmäßig [in Deutschland] lebenden Ausländern allein abgeschoben
-1	-1	19509	lebenden	1887	66%	in hier|Deutschland lebenden [...] Ausländer
-2	-2	13649	**Integration** lebender	34	100%	die Integration hier|in Österreich lebender Ausländer
-2	-2	13649	Integration Aussiedlern	24	75%	die|zur Integration von Ausländern und Aussiedlern
-2	-2	13649	Integration	1933	62%	die Integration von Ausländern
-2	-2	13225	**Ausländerinnen** erleichterte	17	52%	erleichterte Einbürgerung junger Ausländerinnen und Ausländer
-2	-2	13225	Ausländerinnen Schweiz	45	60%	Ausländerinnen und Ausländer in|die der Schweiz
-2	-2	13225	Ausländerinnen Stimm	15	53%	Ausländerinnen und Ausländern ... das Stimm und Wahlrecht
-2	-2	13225	Ausländerinnen	950	67%	Ausländerinnen [und] Ausländer
-4	5	7413	**Deutschland** lebende	228	98%	in Deutschland [...] lebende [...] Ausländer
-4	5	7413	Deutschland geborene	123	57%	in|In Deutschland geborene Kinder von Ausländern die
-4	5	7413	Deutschland geborenen	86	82%	in Deutschland [...] geborenen [Kinder|Kindern von] Ausländern die ...
-4	5	7413	Deutschland	3688	42%	Ausländer [... in] Deutschland
-5	4	6904	**illegal** eingereiste	130	96%	illegal [...] eingereiste [...] Ausländer
-5	4	6904	illegal eingereisten	41	63%	von illegal [...] eingereisten [...] Ausländern
-5	4	6904	illegal beschäftigte	72	90%	illegal [...] beschäftigte Ausländer
-5	4	6904	illegal	1107	58%	illegal [in ...] Ausländer
-1	-1	6673	**lebende** hier legal	5	100%	legal hier lebende Ausländer
-1	-1	6673	lebende hier	153	96%	für hier [...] lebende [...] Ausländer
-1	-1	6673	lebende legal	36	100%	legal [in Österreich] lebende Ausländer
-1	-1	6673	lebende	744	97%	in hier|Deutschland lebende [...] Ausländer
-1	-1	6022	**viele** leben	143	62%	zu viele [...] Ausländer [...] leben
-1	-1	6022	viele lebten	29	62%	es|Deutschland lebten zu viele [...] Ausländer in
-1	-1	6022	viele wohnen	49	67%	in dem viele [...] Ausländer [...] wohnen
-1	-1	6022	viele	1867	84%	viele [...] Ausländer

Abb. 2.2: Kookkurrenzprofil von „Ausländer“ (Ausschnitt), Kookkurrenzdatenbank CCDB (Belica 2001 ff.).

Deutsche Sprache zeigt (vgl. Abb. 2.3; Kleiner 2011 ff.). Die Kartendarstellung als diagrammatische Grundfigur fügt zu sprachlichen Äußerungen eine weitere Dimension, nämlich eine geografische, hinzu.

Die diagrammatische Grundfigur der *Partitur* entstand in Ansätzen bereits im neunten Jahrhundert, um die Polyphonie von Musik sichtbar zu machen (Sachs & Röder 1989), also die Gleichzeitigkeit von Stimmen. Offensichtlich hat diese Idee der Notation in den 1970er Jahren Eingang in die Gesprächsanalyse gefunden (Sacks, Schegloff & Jefferson 1974; Redder 2001), wo sie ein wichtiges Element war, um gesprochene Sprache überhaupt unter einer neuen Perspektive, die den Turn in den Fokus nimmt, zu konstituieren (vgl. für weitere Ausführungen dazu Bubenhofer 2018c). In der Korpuslinguistik ist diese Grundidee jedoch viel unscheinbarer auch in Annotationssystemen enthalten: Eine Auszeichungssprache wie XML fügt genauso Ebenen oder „Stimmen“ zu einem

Abb. 2.3: Realisierung des auslautenden Konsonanten <ig> im abgeleiteten Abdjektivabstraktum *Notwendigkeit*, http://prowiki.ids-mannheim.de/bin/view/AADG/IgKeit (letzter Zugriff: 6. 11. 2017), (Kleiner 2011 ff.).

Primärtext hinzu, wie das die Partitur macht.[4] Auch die Darstellung von Korpora als „vertikalisierter“ Text, indem pro Zeile ein Token und durch Tabulator oder ein anderes Trennzeichen separiert weitere Informationen zum Token (Wortart, Grundform etc.) als „Spalten“ hinzugefügt werden, nimmt auf die Grundfigur der Partitur Bezug. Die diagrammatische Grundfigur der Partitur erlaubt es also, Text zwar als sequenziell, jedoch als komplexes Mehr-Ebenen-Phänomen darzustellen und fügt zu dem Text weitere Dimensionen hinzu.

Die Grundfigur der *Vektoren* ist in der Korpuslinguistik besonders relevant: Um die gängigen Formen der Darstellung von Linien-, Balken-, Punktdiagrammen etc. zu ermöglichen, ist vorher die Transformation von Sprache in einen Vektorraum notwendig. Ausgehend von einer Index-Liste von Belegen, werden diese in eine Tabelle von Vektoren überführt, wo beispielsweise Frequenzen nach Jahren geordnet werden, um ein Liniendiagramm zur diachronen Entwicklung eines Phänomens zu erstellen (vgl. dazu etwa die Arbeiten des Lexik-Projekts „Empirische Methoden“ am IDS, z. B. zu diachronem Korpuswandel, Koplenig 2017). Eine Erweiterung ist die Idee, komplexere Vektoren zu erstellen, etwa Kollokationsprofile: Diese werden als Vektor repräsentiert, der die Frequenzen des Kovorkommens des Basislemmas mit den Kollokatoren enthält. Dadurch wird die Semantik (wenn man einem Begriff der distributionellen Semantik folgt) als Vektor codiert und damit mit anderen Vektoren verrechenbar: Es lassen sich Abstände, Homologien etc. berechnen, wie dies etwa bei Verfahren des Word Embeddings geschieht (Mikolov et al. 2013).[5] In gleicher Weise, aber auf Ebene von Texten, wird beim Topic Modelling verfahren (Graham, Weingart & Milligan 2012). Die diagrammatische Grundfigur der Transformation sprachlicher Daten in Vektoren ist der eigentliche Schritt der „Verdatung“ von Sprache, um sie mathematischen Operationen zugänglich zu machen (Bubenhofer & Rothenhäusler 2016: 63) und damit am Beispiel der distributionellen Semantik eine Wortbedeutung zu modellieren, die eine „*fiktive*, mathematisch aus Beobachtungsdaten erzeugte Entität“ (Bender & Marrinan 2014: 197) darstellt.

Graphen in gerichteter oder ungerichteter Form gehören ebenfalls zu sehr alten Formen von Diagrammen, die etwa in Form von Stammbäumen bereits vor über 2000 Jahren gezeichnet wurden (Kruja et al. 2002; Lima 2014). In der

4 XML leistet darüber hinaus aber natürlich noch mehr, indem beispielsweise auch eine Hierarchisierung der Ebenen zueinander abbildbar ist. Vgl. aber Bański (2010) gerade für die Probleme von XML zur Annotation von Text.

5 Vgl. für eine Kombination von Word Embeddings und diachronem Vergleich die Anwendung „DiaViz“ von Peter Fankhauser und Marc Kupietz (IDS): http://corpora.ids-mannheim.de/diaviz/dereko.html (letzter Zugriff: 6. 11. 2017)

Korpuslinguistik werden Graphen insbesondere für die Visualisierung von Kollokationen verwendet (Bubenhofer 2018b; Brezina, McEnery & Wattam 2015), so auch in der CCDB des IDS (Belica 2001 ff.) oder im Rahmen des Projekts „Usuelle Wortverbindungen" (Steyer 2013). Aus diagrammatischer Sicht zeichnet sich die Graphdarstellung dadurch aus, dass die strukturelle Information darüber, welche Knoten Verbindungen zueinander aufweisen, von der darstellerischen entkoppelt ist: Die Definition der Knoten-Kanten-Beziehungen muss für die Darstellung als Netzwerkgraph in ein Layout überführt werden, also meist ein algorithmisch definierbares Prinzip, nachdem die Knoten angeordnet werden. Im Kontext der Netzwerkanalysen werden oft „force-directed"-Layoutmodi verwendet, die die Knoten nach den physikalischen Prinzipien der Anziehung und Abstoßung in einem energetischen Optimum platzieren. Knoten, die viele Verbindungen untereinander haben, tendieren dann dazu, nahe beieinander zu stehen. Ein solches Bild ist semantisch (bei Kollokationen), hermeneutisch (z. B. bei typischen Wortverbindungen in Geschichten o. ä.) etc. interpretierbar.

1.3 Präsentation vs. Exploration

Eine wichtige Unterscheidung von Visualisierungstypen generell, die quer zu den oben beschriebenen Typen steht, ist jene von Präsentations- und Explorationsgrafiken (Chen, Härdle & Unwin 2008: 4–5; Schumann & Müller 1999: 5). Erstere dienen dazu, der Forscherin oder dem Forscher bereits bekannte Erkenntnisse in einem Diagramm darzustellen, etwa um sie besser lesbar zu machen. Explorative Visualisierungsmethoden hingegen werden im Analyseprozess eingesetzt, um die Daten überhaupt interpretierbar zu machen. Das Paradigma der „Visual Analytics" (Keim et al. 2010; Chen, Härdle & Unwin 2008) nutzt solche Techniken und verbindet so quantitativ-maschinelle mit qualitativ-interpretierenden Analysemethoden:

> Visualisation becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective, distinct capabilities for the most effective results (Keim et al. 2010: 14).

Im Anschluss an die diagrammatischen Überlegungen in Abschnitt 1.1 lässt sich sagen, dass explorative Visualisierungen Diagramme mit ausgeprägter Operationalität sind: Im besten Fall nutzen explorative Visualisierungen ein Diagrammschema, mit dem die Daten so dargestellt werden, dass dank der Darstellung neue Erkenntnisse gewonnen werden können. Dies ist beispielsweise der Fall, wenn Kollokationen als Netz dargestellt werden, wie in Ab-

schnitt 1.2 zur diagrammatischen Grundfigur Graph bereits ausgeführt worden ist (Pfeffer 2010; Kruja et al. 2002; Chen, Härdle & Unwin 2008: 109; Bubenhofer 2018b).

Die in einem Netzwerkgraph sichtbar gewordenen Cluster von Knoten sind nun eine Erkenntnis, die aufgrund der Listen alleine wohl nicht hätte gewonnen werden können (vor allem bei großen Datenmengen). Das Beispiel macht jedoch auch deutlich, dass nicht nur die grafische Umsetzung alleine den Mehrwert für die Analyse ermöglichte, sondern Darstellungsprinzipien genereller Natur: das Netz, die physikalischen Prinzipien, die statistische Berechnung etc.

Mit einem weiten Diagramm-Begriff könnte man argumentieren, dass diese generellen Darstellungs- und Ordnungsprinzipien (wie Listen, Matrizen, Vektorräume etc.) zum grafischen Diagramm dazu gehören und Operationalität im umfassenden Sinn ermöglichen. Es ist ein Spezifikum von Digitalität, dass die verschiedenen Formen der Materialisierung von Daten (als Bildschirmbild, Ausdruck etc.) mit den immateriellen Repräsentationen und Manipulationen von Daten einhergehen und darauf aufbauen (Bubenhofer 2018c). Diagrammatische Operationen umfassen deshalb mehr als Interaktivität in einem Diagramm; dazu gehören auch die Operationen der immateriellen Repräsentation und Manipulation von Daten.

Bei einem engen Diagramm-Begriff würde man die digitalen Operationen nicht als diagrammhaft verstehen, sondern erst die grafische Repräsentation des Diagramms. Doch auch dann ist augenfällig, dass insbesondere beim digitalen Diagramm die Verbindung zu den Daten so eng ist, dass eine scharfe Trennung zwischen Diagramm und Nicht-Diagramm schwierig ist.

Präsentationsgrafiken sind oft weniger offensichtlich operational. Ein Rest an Operationalität ist aber auch da vorhanden, da sie genauso als „Denkzeuge" (Krämer 2016: 83) fungieren. Die Darstellung eines Ausschnitts aus einer KWiC-Liste in einer Publikation legt eine bestimmte (von der Erstellerin erwünschte) Interpretation nahe, verhindert aber nicht, dass der Rezipient damit noch andere Erkenntnisse gewinnt (auch wenn sie unvollständig sein mögen), die der Erstellerin entgangen sind oder nicht im Fokus standen.

2 Explorative Visualisierungen: Anwendungsbeispiele

Die bisherigen Ausführungen waren theoretischer Natur. Im Folgenden möchte ich in aller Kürze von zwei Experimenten berichten, bei denen wir visuelle

Analysemethoden einsetzten, um Korpusdaten besser zu verstehen. Im vorliegenden Beitrag geht es mir um die diagrammatischen Komponenten und die damit zusammenhängenden Reflexionen.

2.1 Geokollokationen

Die Berechnung und Darstellung von „Geokollokationen" zielt darauf ab, diskursiv geprägte Konstruktionen von Welt zu beschreiben (Bubenhofer 2014; Bubenhofer et al. 2017). In Diskursen entstehen bestimmte Assoziationen zu Orten und Regionen: *Schweiz – Banken*, *Schokolade*, *Steuerhinterziehung*; *Griechenland – Finanzkrise*, *Urlaub*, *Flüchtlingskrise* etc. Das Ziel des Experiments ist es, datengeleitet aus Korpora solche Assoziationen abzuleiten.

Als Datenbasis dienten uns verschiedene Quellen (Presse, Bundestagdebatten, Webdiskussionsforen etc.). Die im Folgenden präsentierten Daten beruhen auf einem Korpus von Artikeln des Magazins *Der Spiegel* und der Wochenzeitung *Die Zeit* von 1946 bis 2016 (611 Mio. Tokens). Die Daten wurden mit dem TreeTagger (Schmid 1994, 1995) und der Standardbibliothek für Deutsch lemmatisiert und mit Wortarten nach dem Stuttgart-Tübingen-Tagset (Schiller, Teufel & Thielen 1995) annotiert. Zusätzlich wurde der Stanford Named Entity Recognizer (NER) (Finkel, Grenager & Manning 2005) in einer fürs Deutsche adaptierten Version (Faruqui & Padó 2010) auf die Daten angewandt.

Ausgehend von den vom NER-Tagger als Toponyme erkannten Lexeme wurden die zu dieser Basis statistisch auffälligen Kollokatoren im gleichen Satz mittels Log Likelihood Signifikanztest (Evert 2009) berechnet. Die daraus resultierenden Listen sind lang und umfassen Toponyme und damit kollokierende Lexeme. Abbildung 2.4 zeigt einen Ausschnitt aus einer solchen Liste, die jedoch bereits mit weiteren Informationen, darunter Georeferenzen angereichert ist.

Um die Unübersichtlichkeit der Liste zu beheben, ist es naheliegend, die Daten auf eine Karte zu projizieren. Dafür ist die Georeferenzierung – also die Anreicherung der Toponyme mit Geokoordinaten – notwendig, der wiederum oft eine Disambiguierung vorausgehen muss. Für die Disambiguierung verwenden wir den „Cartographic Location and Vicinity Indexer" (CLAVIN),[6] der Ortskandidaten nach Populationsgröße ordnet (*Berlin* in Deutschland vor *Berlin* in den USA) und danach den Textkontext berücksichtigt und Orte bevorzugt, die nahe beieinander liegen.

---

6 Vgl. https://github.com/Berico-Technologies/CLAVIN/ (letzter Zugriff: 6. 11. 2017).

```
name lon lat type country cow_code freq sig word pos
"Yunnan" 102 25 state cn 31 0.0001 "Provinz" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 27 0.0001 "töten" VVPP
"Sinai" -78.1885830257228 -2.08513105 administrative ec 32 0.0001 "werden" VAPP
"Sinai" -78.1885830257228 -2.08513105 administrative ec 36 0.0001 "Al-Arisch" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 23 0.0001 "Islamisten" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 54 0.0001 "Halbinsel" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 24 0.0001 "Mann" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 30 0.0001 "öffentlich" ADJA
"Sinai" -78.1885830257228 -2.08513105 administrative ec 21 0.0001 "Al-Arisch" ADJD
"Sinai" -78.1885830257228 -2.08513105 administrative ec 20 0.0001 "Demonstrant" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 39 0.0001 "Norden" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 24 0.0001 "Polizist" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 25 0.0001 "Stadt" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 29 0.0001 "ägyptisch" ADJA
"Sinai" -78.1885830257228 -2.08513105 administrative ec 26 0.0001 "Extremist" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 30 0.0001 "Gebäude" NN
"Sinai" -78.1885830257228 -2.08513105 administrative ec 36 0.0001 "bewaffnet" ADJA
"Dominikanische Republik" -70.3012705 19.094175 administrative do 68 0.0001 "Republik" NN
"Dominikanische Republik" -70.3012705 19.094175 administrative do 50 0.0001 "Dominikanische" NN
"Katar" 51.2295295 25.3336984 administrative qa 63 0.0001 "Präsident" NN
"Katar" 51.2295295 25.3336984 administrative qa 56 0.0001 "Todesfall" NN
"Katar" 51.2295295 25.3336984 administrative qa 31 0.0001 "Fußball-Bund" NN
"Katar" 51.2295295 25.3336984 administrative qa 49 0.0001 "Menschenrecht" NN
"Katar" 51.2295295 25.3336984 administrative qa 22 0.0001 "Temperatur" NN
"Katar" 51.2295295 25.3336984 administrative qa 50 0.0001 "Turnier" NN
"Katar" 51.2295295 25.3336984 administrative qa 29 0.0001 "geplant" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 24 0.0001 "alarmierend" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 74 0.0001 "international" ADJA
"Katar" 51.2295295 25.3336984 administrative qa 24 0.0001 "Situation" NN
```

Abb. 2.4: Ausschnitt aus einer Liste von Geokollokationen: In der ersten Spalte das Toponym, in den letzten beiden Spalten der Kollokator mit Wortartklasse.

Nun ist es möglich, die Daten auf einer Karte darzustellen. Dafür verwenden wir Javascript und insbesondere die Bibliothek „D3" (Bostock, Ogievetsky & Heer 2011). Abbildung 2.5 zeigt einen Ausschnitt aus der Kartendarstellung: Grundsätzlich werden Orte, an denen Kollokatoren vorhanden sind, als Punkte dargestellt, wobei die Punktgröße die Anzahl der Kollokatoren repräsentiert. Je nach Einstellung im Kontrollfeld der Karte werden anstelle der Punkte immer (oder nur bei genügend vorhandenem Platz) die Kollokatoren direkt als Text angezeigt.

Die Geokollokationen-Visualisierung ist ein exploratives Werkzeug und umfasst deshalb auch ein Kontrollfeld mit verschiedenen Einstellmöglichkeiten, um die Daten zu filtern: Auswahl des Datensatzes (des Korpus), Setzen von Schwellwerten, Restriktion auf bestimmte Wortartklassen und eine Einschränkung auf Kollokatoren, die auf einen regulären Ausdruck passen. So zeigt Abbildung 2.6 Orte und Staaten, die im Zusammenhang mit *Flucht*, *Flüchtling* oder *Migration* genannt werden.[7]

7 Gesucht wurde nach Kollokatoren, die auf den regulären Ausdruck `.*([Ff]l[uü]cht|[Mm]igra).*` passen.

Abb. 2.5: Ausschnitt aus der Kartendarstellung der Geokollokationen: Anzeige von Orten mit Kollokatoren als Punkte und mit Text bei genügend Platz.

Abb. 2.6: Ausschnitt aus der Kartendarstellung der Geokollokationen: Restriktion auf Kollokatoren, die die Zeichenkette *flucht*/*flücht* oder *migra* enthalten, also zu den Themen *Flüchtlinge* und *Migration* – Punktgröße repräsentiert die Anzahl der Kollokatoren.

Abb. 2.7: Ausschnitt aus der Kartendarstellung der Geokollokationen: Restriktion auf Kollokatoren, die die Zeichenkette *flucht*/*flücht* oder *migra* enthalten, also zu den Themen *Flüchtlinge* und *Migration* – Korpus Zeit/Spiegel 2010–2016, Anzeige der Kollokatoren.

Die Kartendarstellung erlaubt es nun in der Folge, geografische Zusammenhänge zwischen den Toponymen und ihren Assoziationen zu entdecken. So wird sichtbar, dass bestimmte Kollokatoren sehr global verwendet werden, z. B.: *Stadt*, *Land* oder *Jahr*. Andere hingegen sind spezifisch für bestimmte Regionen, die eine Gemeinsamkeit aufweisen, wie beispielsweise *Menschenrecht* oder *Flüchtling*. Andere sind sehr ortsspezifisch wie z. B. *chinesisch* oder *Obama*.

Schränkt man die Anzeige der Kollokatoren ein, um einen thematisch definierten Diskurs zu untersuchen, lassen sich die Nuancen der Berichterstattung dazu und die damit konstruierten geografischen Assoziationen entdecken. In Abbildung 2.7 sind die relevanten Kollokatoren im Bereich Flucht/Migration mit der selben Einschränkung wie in Abbildung 2.6 im Raum Deutschland, Balkan, Türkei, Naher Osten sichtbar. Auffallend sind die vielen Derivationen (hauptsächlich Nominalkomposita) von den Lexemen *Flüchtling*, *Flucht* und *Migration* wie etwa *Flüchtlingslager*, *Zuflucht*, *Bootsflüchtling*, *Flüchtlingswelle*, *Flüchtlingszahl* etc. Allerdings werden diese Derivationen hauptsächlich im

Zusammenhang mit Deutschland und Europa (gemeint ist dabei meist die Europäische Union) genannt, nicht mit den Ursprungsländern der Flucht. Dort gibt es generell weniger Derivationen. Dies deutet darauf hin, dass im Diskurs Migration primär als innenpolitisches Thema konstruiert wird – in der deutschen Presse als deutsches und EU-Problem – und nicht als Problem der Ursprungsländer oder der Länder, die außerhalb Deutschlands als Transitländer davon betroffen sind.

Dank der Korpusdaten von *Der Spiegel* und *Die Zeit*, die die ganze Nachkriegszeit abdecken, können auch diachrone Veränderungen analysiert werden. Wechselt man etwa bei der gleichen Einschränkung der Kollokatoren auf den Migrationsdiskurs die Datengrundlage und wählt anstelle des Zeitraums 2010 bis 2016 die Nachkriegsjahre, werden die Unterschiede alleine anhand der damit assoziierten Regionen sichtbar. Anstelle Afrikas wird der ganze amerikanische Kontinent damit verknüpft, was natürlich daran liegt, dass aus deutscher Perspektive Migration in der Nachkriegszeit auch ein Emigrationsthema war.[8]

Die Darstellung der Geokollokationen auf einer Karte ist naheliegend und unterstützt eine Denkfigur, die in unseren Köpfen wahrscheinlich automatisch anspringt, wenn wir Toponyme lesen: die geografische Verortung. Diese Denkfigur ist geprägt von den uns bekannten Kartenbildern. Karten sind jedoch immer zweidimensionale Projektionen der kugelartigen Welt, die immer verzerrt ist und beispielsweise bei der uns vertrauten Mercator-Projektion die Länder am Äquator im Vergleich zu den davon entfernten Ländern viel kleiner darstellt (Glasze 2009; Smith 1992). Zudem stellt sich das Problem, dass die Größe der abgebildeten Länder geografisch vorbestimmt ist und nicht zwingend auch die diskursive Bedeutung widerspiegelt. Daher ist es gerade interessant, unterschiedliche Visualisierungslösungen auszuprobieren – und damit mit dem Diagramm ein anderes System der Relationen zu konstruieren.

Dafür implementierten wir eine sog. Dorling-Ansicht (Dorling 1993), bei der die geografischen Entitäten (z. B. Staaten) als Punkte dargestellt werden, deren Größe eine Datenvariable repräsentiert. Die Positionierung der Punkte folgt zwar einer geografischen Ordnung, geht aber zwangsweise Kompromisse ein, um Überlappungen zu vermeiden. Die Größe der Punkte repräsentiert in unserem Fall die Anzahl signifikanter Kollokatoren zum jeweiligen Staat.

Diese Darstellung kombinierten wir mit einem Sankey-Diagramm (Sankey 1896), das Flüsse zwischen Entitäten in Form von unterschiedlich breiten Linien darstellt. In unserem Fall nutzten wir diese Darstellung, um eine (von der Benutzerin/dem Benutzer) ausgewählte Zahl von Kollokatoren unterhalb

8 Siehe für weiterführende Ausführungen dazu auch Bubenhofer et al. (2017).

Abb. 2.8: Kombinierte Darstellung der Karte als Dorling- und Sankey-Diagramm. Ausgewählt ist der Kollokator *Frühling*, der auf die Länder, die mit dem *Arabischen Frühling* assoziiert werden, verweist.

der Karte darzustellen und anzuzeigen, mit welchen Ländern diese kollokieren. Abbildung 2.8 zeigt diese Ansicht.

Dieser Versuch, von der geografischen Darstellung zu abstrahieren, kommt einer Darstellung der diskursiv geprägten Weltsicht näher als die konventionelle Karte. Sie greift die Diskussionen der kritischen Kartographie (Glasze 2009) auf und deutet an, dass es immer eine Vielzahl von Möglichkeiten gibt, das ikonische Ähnlichkeitsverhältnis zwischen Denotat und Diagramm zu konstruieren.

2.2 Narrative

Im Fall der Geokollokationen (in Abschnitt 2.1) ist die Darstellung der Daten auf einer Karte eine naheliegende Lösung, die dabei hilft, Bezüge zwischen den Datenpunkten sichtbar zu machen. Beim zweiten Beispiel für explorative Visualisierung geht es darum, überhaupt erst adäquate Formen der Visualisierung zu finden, um die Daten analysierbar zu machen. Es geht darum, typische narrative Muster in seriellen Alltagserzählungen (Ehlich 1980; Gülich 1980; Quasthoff 1980) zu finden. So gibt es beispielsweise Webdiskussionsforen, in denen Frauen die Geburt ihrer Kinder erzählen. Für solche Erzählungen wird meist ein formaler Rahmen innerhalb allgemeiner Foren zu Elternschaft bereit-

Tab. 2.1: Korpus „Geburtsberichte“.

	# Wörter	# Texte
http://www.urbia.de/	7.364.108	8.808
http://www.babyforum.de/	2.089.936	1.824
http://www.parents.at/	1.199.174	1.647
https://www.swissmomforum.ch/	1.156.193	919
http://www.eltern.de/	438.017	716
http://www.umstandsforum.de/	289.807	568
Total	**12.537.235**	**14.482**

gestellt, meist ein Unterforum mit dem Titel „Geburtsberichte“ o. ä. Dort schreiben die Mütter dann in Form von Postings ihre Erlebnisse und andere Leserinnen[9] kommentieren diese Initialpostings. Auffallend ist eine deutliche Serialität der Erzählungen: Obwohl das Geburtserlebnis individuell einmalig ist, sind es Geburten überhaupt nicht. Und auch die Erzählungen darüber folgen offensichtlich bestimmten Folien, die soziokulturell verankert sind.

Ich werde an dieser Stelle nur kurz die Eckdaten der Studie skizzieren und hauptsächlich die diagrammatischen Aspekte problematisieren. Im Detail wird die Studie in Bubenhofer (2018a) vorgestellt.[10]

Ziel der Studie ist es, solche „narrativen Muster“ der Geburtserzählung (in den erwähnten Foren) datengeleitet zu berechnen. Dafür stellte ich ein Korpus von über 14.000 Geburtsberichten (12 Mio. Tokens) aus sechs Foren im deutschsprachigen Raum zusammen (vgl. Tab. 2.1).

Die Korpusdaten wurden mit Hilfe des TreeTaggers (Schmid 1994) lemmatisiert und mit dem Stuttgart-Tübingen-Tagset (Schiller, Teufel & Thielen 1995) mit Wortartklassen annotiert. Um typische Formulierungsmuster in den Berichten zu finden, verwendeten wir den bereits an verschiedenen Daten erprobten Ansatz, typische n-Gramme zu extrahieren (Bubenhofer 2017). Dazu benutzten wir ein Referenzkorpus von Presseartikeln aus *Die Zeit* und *Der Spiegel* – das gleiche Korpus wie in Abschnitt 2.1 angegeben, allerdings nur den Zeitraum von 2010 bis 2016 – und berechneten in jedem Korpus alle auftretenden Wort-n-Gramme mit $n > 5$. Für die Berechnung berücksichtigten wir jedoch jeweils die Grundformen (Lemmata) anstelle der Wortformen. Zudem durfte sich ein n-Gramm nicht über einen Satz hinaus erstrecken. Anschließend verglichen

9 Es scheint sich fast immer um Frauen zu handeln.

10 Vgl. für eine breite Einführung in die Erzählforschung zu Geburtsberichten Colloseus (2016) und für eine kleinere linguistische Studie dazu Barbieri et al. (2012).

wir die Frequenzen der n-Gramme in den beiden Korpora und führten einen Log Likelihood-Signifikanztest durch, um die n-Gramme, die typisch für die Geburtsberichte sind, zu extrahieren (inkl. jener, die nur im Geburtsberichte-Korpus vorkommen).

Soweit verfolgt die Methode einen sog. „Bag of Words"-Ansatz, bei dem zwar nicht Einzellexeme, sondern n-Gramme, jedoch ohne Rücksicht auf ihre

Abfolgen und Positionen in den Geschichten, erfasst werden. Um narrative Muster aufdecken zu können, ist es jedoch notwendig, typische Sequenzen von n-Grammen zu finden. Daher nutzten wir folgende zwei Strategien:

– Zu jedem n-Gramm-Token wurde die relative Position in der Geschichte (zwischen 0 – Anfang – und 1 – Ende der Geschichte) erfasst, so dass pro n-Gramm-Type der Mittelwert und die Standardabweichung der Positionen berechnet werden kann.
– Weiter berechneten wir für jeden n-Gramm-Type die damit links und rechts kollokierenden n-Gramm-Types, in Anlehnung an den in Bubenhofer, Müller & Scharloth (2013) beschriebenen Ansatz.[11]

Abbildung 2.9 zeigt im Überblick die für die Geburtsberichte signifikanten n-Gramme und ihre relativen Positionen in den Geschichten in Korrelation zur Standardabweichung der Positionen. Dabei ist eine deutliche Korrelation zwischen Position und Variation ersichtlich: Die n-Gramme am Anfang und am Ende der Geschichten sind in ihren Positionen relativ stabil, während in der Mitte der Geschichten mehr Variation in der Position zu beobachten ist. Das sehr häufige n-Gramm *das Licht der Welt erblickt* kommt zwar im Mittel tatsächlich auch etwa in der Mitte der Geschichte vor, die Standardabweichung ist jedoch relativ hoch, da es oft auch eher am Anfang oder Ende der Geschichten auftritt. Ein n-Gramm wie *der eigentliche ET war der*[12] steht jedoch in fast allen Geschichten jeweils am Anfang.

Es gibt nun verschiedene Kriterien, nach denen die n-Gramme angeordnet werden können. Die relative Position in der Geschichte ist dabei das wichtigste Kriterium. Frequenz, Standardabweichung der Position aber auch Ähnlichkeit der n-Gramme zueinander sind aber weitere wichtige Kriterien. Um eine flexible Exploration der Daten zu ermöglichen, experimentierten wir mit interaktiven, dreidimensionalen Darstellungen (vgl. das Bildschirmfoto in Abb. 2.10). Dreidimensionale Diagramme bedürfen der Interaktion und können

11 Die genaue Implementierung ist in Affolter (2016) beschrieben.
12 „ET" steht für „errechneter Termin".

Abb. 2.9: Positionen der vorhandenen n-Gramme in den Geburtsberichten in Korrelation zur Standardabweichung der Position (Abb. aus Bubenhofer 2018a).

auf Papier nur unzureichend dargestellt werden, weswegen ich auf die Online-Version verweise.[13]

Der Vorteil von interaktiven, dreidimensionalen Darstellungen ist, dass je nach Bedarf verschiedene Perspektiven auf die Daten möglich sind. Bei der Vorderansicht (x-Achse von links nach rechts, y-Achse von unten nach oben) wie in Abbildung 2.10 wird die Korrelation von Position und Frequenz fokussiert, eine Draufsicht (z statt y-Achse) zeigt die verschiedenen Varianten ähnlicher n-Gramme. Mit den zahlreichen „Zwischenperspektiven" können die Ordnungskriterien beliebig gewichtet werden.

13 Vgl. https://www.bubenhofer.com/sprechtakel/2017/02/19/die-serielle-singularitaet-vierzehntausend-geburtsgeschichten/ (letzter Zugriff: 6. 11. 2017).

Abb. 2.10: Dreidimensionale Darstellung der Gebuchtsberichte-n-Gramme – x-Achse: Relative Position; y-Achse: Frequenz; z-Achse: Ähnlichkeit (Abb. aus Bubenhofer 2018a).

In den Darstellungen in Abbildung 2.9 und Abbildung 2.10 sind die Kollokationsinformationen, also die Angaben darüber, welche n-Gramm-Sequenzen typischerweise vorkommen, nicht dargestellt. Um diese Informationen darstellbar zu machen, entwickelten wir das Visual-Analytics-Tool „NarrViz“ (implementiert von Katrin Affolter, vgl. Affolter 2016). Abbildung 2.11 zeigt die Oberfläche mit dem geladenen Datensatz der Geburtsberichte. Es handelt sich um eine in Javascript realisierte Webbrowser-Anwendung.[14]

Auch in NarrViz repräsentieren die Knoten die für die Geburtsberichte signifikanten n-Gramme. Sie sind in Spalten von links nach rechts angeordnet, jede Spalte (mit Farbcodierung) entspricht ein relativer Positionsbereich in den Geschichten. Knoten links stehen also am Anfang der Geschichten, jene rechts am Ende. Die Größe der Knoten steht für die Frequenz des n-Gramms. Signifikante Assoziationen zwischen den n-Grammen sind durch die Kanten dargestellt.

Die Visualisierung ist interaktiv: Bei Berührung eines Knotens werden verschiedene Informationen zum n-Gramm angezeigt. Abbildung 2.12 zeigt den ausgewählten Knoten zum n-Gramm (in lemmatisierter Form) *ich haben die Gefühl*, mit der Angabe der häufigsten Exemplare des abstrakten n-Gramms (*Ich hatte das Gefühl*, als häufigste Form). Ebenso angegeben sind statistische

14 Die Anwendung kann unter https://pub.cl.uzh.ch/projects/visuallinguistics/NarrViz/ (letzter Zugriff: 6. 11. 2017) ausprobiert werden.

Abb. 2.11: Visual-Analytics-Tool „NarrViz“, Oberfläche mit Datensatz Geburtsberichte.

Abb. 2.12: Visual-Analytics-Tool „NarrViz“, Informationen zu einem n-Gramm.

Maße zum Knoten: Frequenz, Signifikanzniveau des n-Gramms (im Vergleich zum Referenzkorpus) und die Gradzentralität, also die Anzahl eintreffender (in-degree, δ^-) und ausgehender (out-degree, δ^+) Verbindungen. Darunter ist die Verteilung des n-Gramms über die Positionen in den Geschichten als Boxplot dargestellt: Bei der Annahme von zehn Teilen befindet sich das n-Gramm im Mittel im sechsten Teil, wobei die Hälfte der Menge zwischen dem vierten und siebten Teil streut. Ausreißer gibt es aber in allen Teilen, außer dem ersten.

Unterhalb des Boxplots zur Verteilung ist der wahrscheinlichste Pfad, also die wahrscheinlichste Sequenz aufgeführt, in der das n-Gramm vorkommt: *auf den Weg machen*, → *so stark, dass ich* → *Ich hatte das Gefühl*, → *Für mich war es eine*.[15]

Ist einer der Knoten ausgewählt, wird dieser Knoten zusammen mit seinen Verbindungen zu den anderen Knoten in der Netzwerkvisualisierung entsprechend hervorgehoben. Daneben gibt es umfangreiche Möglichkeiten, die Daten zu filtern und die Darstellung zu beeinflussen. Dies ist notwendig, da die sich ergebenden Graphen je nach Datengrundlage unterschiedlich komplex werden und bei sehr komplexen Daten gefiltert werden muss, um vernünftig arbeiten zu können.[16]

Nicht alle n-Gramme sind bezüglich ihrer Position stabil. Die Darstellung als Knoten an einer bestimmten Position ist für solche n-Gramme deshalb irreführend. Daher gibt es in NarrViz die Möglichkeit, Knoten ab einem festlegbaren Schwellwert der Standardabweichung von der mittleren Position als Balken darzustellen, deren Länge das obere und untere Quartil des Streubereichs und deren Höhe die Frequenz darstellt (vgl. Abb. 2.13). Damit ergibt sich im Überblick das Bild von n-Grammen, die sozusagen die Basis der Geschichten darstellen und an verschiedenen Positionen auftauchen. Davon heben sich dann die positionsspezifischen n-Gramme ab.

Für die linguistische Interpretation bietet es sich an, linguistische Erzähltheorien heranzuziehen, etwa jene von Labov & Waletzky (1973). Die gefundenen n-Gramme lassen sich relativ gut den dort vorgeschlagenen Erzählfunktionen zuordnen:

– **Orientierung**: *der (eigentliche) ET war der*; *gegen 22:30 Uhr sind wir*; *rief ich meinen Mann an* etc.

---

15 Der besseren Lesbarkeit wegen werden neben der lemmatisierten, abstrakteren Form (für die die Angaben gelten) auch noch jeweils die häufigsten Wortform-Varianten angegeben.

16 Dies widerspricht natürlich dem Paradigma der Visual Analytics, auch komplexe Daten visualisieren zu können. Die Visualisierungslösung ist deswegen auch nicht perfekt und taugt nur für Daten bis zu einem gewissen Komplexitätsgrad. Dazu kommen auch Performance-Probleme, die durch Optimierung der Implementierung behoben werden müssten.

Abb. 2.13: Visual-Analytics-Tool „NarrViz“, Darstellung von stabilen (Punkte, oben) und variablen (Balken mit Punkten, unten) n-Grammen.

- **Komplikation**: *wurde ich ans CTG angeschlossen*; *nach langem hin und her*; *eine Wehe nach der anderen*; *wurden die Wehen immer stärker*; *Dann kam die Ärztin und* etc.
- **Evaluation**: *ich war fix und fertig*; *fühlte sich an, als*; *ich fing an zu weinen*; *ich zitterte am ganzen Körper*; *ich hatte in der ganzen Zeit [das Gefühl]* etc.
- **Resolution**: *und dann ging es los*; *ging alles sehr schnell*; *Um 15.00 Uhr war dann*; *um 16:02 Uhr das Licht*; *auf den Bauch gelegt*; *mir auf die Brust gelegt* etc.
- **Coda**: *ich bin so froh, dass*; *alles in allem war es*; *Geburt noch vor sich haben*; *der schönste Moment in meinem* etc.

Gleichzeitig lassen sich aus den Daten aber diese Kategorien auch anreichern mit auffallenden Topoi und Erzählfiguren. So ist für die Geschichten ein Moment der „Akzeleration“ typisch, das als Beginn der Resolution angesehen kann: Am

Punkt der größten Krise kurz vor der eigentlichen Austreibung ist meist ein Moment narrativer Reflexion beobachtbar (typisch für die „Evaluation“ nach Labov & Waletzky 1973), bei dem das Umfeld verschwindet und das Ich der Erzählerin und ihre Gefühle im Vordergrund stehen. Dieser Stillstand wird dann aber mit n-Grammen wie *dann ging alles ganz schnell*; *dann ging alles sehr schnell* in der Erzählung aufgehoben und die Narration akzeleriert. Damit ist narrativ die Geburt mehr oder weniger bewältigt, was in der Realität natürlich nicht zwingend so ist. Die Nachgeburt oder die Wundversorgung im Anschluss sind in den Berichten aber folgerichtig auch oft kein (großes) Thema mehr. Zusammen mit der Beobachtung, dass die Geschichten meist auch gleich nach der Austreibung des Kindes mit n-Grammen wie *um 15.00 Uhr war dann*; *wurde am 29.5.2010 um 02.23 Uhr*; *um 2.24 uhr auf die welt* sozusagen die formelle Geburtsanzeige simulieren, macht die Divergenz zwischen Erlebnis und Narration deutlich: Der genaue Zeitpunkt der Geburt spielte bei der tatsächlichen Geburt wahrscheinlich für die Mutter keine Rolle, wird hinterher jedoch als wichtiger Dreh- und Angelpunkt der Geschichte konstruiert.

2.3 Transformationen und Operationen

Ich kann an dieser Stelle nicht weiter auf die inhaltliche Analyse eingehen[17] und möchte stattdessen einige diagrammatische Überlegungen anstellen. Bei beiden Fallbeispielen (Geokollokationen und Narrative) sind die Visualisierungen eine wichtige Hilfe für die Datenexploration. Sie hängen mit grundsätzlichen Transformationen zusammen, mit denen der Untersuchungsgegenstand erzeugt wurde. Mir scheinen insbesondere vier Grundtransformationen relevant zu sein: Rekontextualisierung, Desequenzialisierung, Dimensionsanreicherung und Rematerialisierung. Was ist damit gemeint?

Jeglicher quantitativer Korpusanalyse eigen ist eine *Rekontextualisierung* von sprachlichen Einheiten: Eine KWiC-Liste ist ein Ensemble von Einzelbelegen, die aus ihren ursprünglichen Kontexten extrahiert und zu einer Liste rekontextualisiert worden sind, die den Untersuchungsgegenstand darstellt. Die Fundstellen werden nicht mehr als Funde innerhalb eines Textes gelesen, sondern als Ensemble aller Fundstellen. Die Einheit des Textes wurde zerstört, um eine Perspektive zu ermöglichen, die nach der Musterhaftigkeit der Verwendung dieser Funde fragt. Bei den Geburtsberichten erzeugten wir durch die datengeleitete Berechnung von typischen n-Grammen den Untersuchungsgegenstand, mit dem die n-Gramme als musterhafte n-Gramme rekontextuali-

17 Vgl. dazu Bubenhofer (2018a).

siert werden: Sie sind ihren ursprünglichen Kontexten entrissen, zeigen dafür die Musterhaftigkeit ihrer Verwendung – erstens weil es sich um Wortsequenzen handelt, die in den Daten gehäuft in diesen Gruppen auftreten, und zweitens weil sie bezüglich ihrer Frequenz auffällig oft in den Geburtsberichten vorkommen. Bei den Geokollokationen hingegen ist das Kollokationsprofil zu jedem Toponym eine Rekontextualisierung von Lexemen im Umfeld des jeweiligen Toponyms, wobei das Kollokationsprofil die Verwendungsweisen kompakt zusammenfasst.

Mit der Rekontextualisierung gehen *Desequenzialisierungen* einher: Bei den Kollokationsprofilen der Geokollokationen sind überhaupt keine Informationen über die syntagmatische Einbettung in den Kontext verfügbar. Die Kombination aus Rekontextualisierung und Desequenzialisierung ist unter korpuslinguistisch-diskursanalytischer Perspektive ein erwünschter Effekt:

> Die dekontextualisierte Darstellung erlaubt es den Forschenden, frei vom ‚hermeneutischen Reflex', der die Lektüre von Texten und Textpassagen bestimmt, kreativ Ideen zu möglichen diskursiven Zusammenhängen einzelner Korpusteile zu entwickeln, die bei einer subjektiven Lektüre möglicherweise verdeckt blieben. (Scholz & Mattissek 2014: 87)

Es ist also einerseits gerade notwendig, vom Einzeltext zu extrahieren (was das Ziel aller quantitativen Analysen ist), andererseits aber auch von der sequenziellen Einbettung der Belege.[18]

Bei den Narrativen ist die Desequenzialisierung kritisch, weil die Abfolge der n-Gramme im Verlauf der Geschichte von großem Interesse ist. Allerdings interessiert uns nicht die einzelne Geschichte, sondern die generalisierte, der musterhafte Ablauf. Daher ist es wichtig, die entsprechenden Daten der typischen Sequenzen zu erheben und die Visualisierung an der Grundfigur der Sequenz auszurichten.

Die dritte relevante Grundfigur ist die *Dimensionsanreicherung*: Bei den Geokollokationen wird durch die Georeferenzierung eine weitere Dimension hinzugefügt mit dem Ziel, eine Visualisierung zu ermöglichen, die die Daten unter einer neuen Perspektive interpretierbar macht. Ich habe auch problematisiert, dass die Kartendarstellung kritisch ist, da es sich um eine kanonische Form der Dimensionsanreicherung handelt, die oft vorschnell als einzig relevante gewählt wird. Daher ist es wichtig, mit unterschiedlichen Anreicherungen zu arbeiten, was wir in Form des Dorling-Diagramms versucht haben.

---

18 Ein Rest an (musterhafter) syntagmatischer Einbettung ist durch die Berechnung der Kollokation natürlich noch vorhanden. Vgl. für eine weiterführende Diskussion auch Bubenhofer (2018c, b).

Bei den Narrativen stellt die Berechnung der typischen Positionen der n-Gramme in den Geschichten die wichtige Dimensionsanreicherung dar, mit der es möglich wird, die typischen narrativen Sequenzen zu finden. Sieht man diese Positionierung der n-Gramme aber auch nur als eine von vielen möglichen Anreicherungen, wird klar, dass auch nach Alternativen gesucht werden muss. Schließlich erzeugen die diagrammatischen Transformationen eine Form von *Rematerialisierung*. Darunter verstehe ich die eigentliche Konstitution des Untersuchungsgegenstands auf einer emergenten Ebene: Ein Kollokationsprofil ist beispielsweise eine statistische Zusammenfassung des Distributionsverhaltens eines Ausdrucks, das durch diagrammatische Transformationen entstanden ist. Bei der Analyse behandeln wir das Profil als Analysematerial, das semantische Lesarten des Lexems darstellt. Ähnlich die sprachlichen Einheiten, die georeferenziert und auf einer Karte dargestellt werden: Sie ergeben einen Gegenstand von Sprache „in situ". Wir können die diagrammatisch manipulierten Daten auf einer emergenten Ebene als einen neuen Gegenstand lesen und interpretieren.

Das Beispiel der Geokollokationen zeigt eine Rematerialisierung als diskursives Bedeutungsgewebe zur Konstruktion von Welt. Dieses Bedeutungsgewebe bewegt sich dabei zwischen stark und weniger stark geografisch verankerten Formen: Bei der Rückbindung auf die Kartenprojektion sind die Abweichungen zwischen diskursivem Weltbild und geografischem besonders deutlich sichtbar. Die Dorling-Visualisierung hingegen spiegelt eher das diskursive Weltbild.

Bei den Narrativen wird durch die Visualisierung die Musterhaftigkeit der typischen Verkettungen gezeigt, also eine Abstrahierung auf zwei Ebenen: Einerseits sind die statistisch auffälligen Sequenzen sichtbar, andererseits aber wiederum eine Abstrahierung der einzelnen auffälligen Sequenzen auf wenige narrative Muster. Diese narrativen Muster erzählen eine Geschichte, wie sie genau so nicht in den Daten zu finden ist, jedoch trotzdem eine passende Typisierung darstellt. Sie lautet z. B. so:

> An diesem Tag hatte ich ... → ..., dass es endlich losgeht → ich hatte das Gefühl, dass ... → Mein Mann und ich waren ... → war mich sicher, dass ... → auf den Weg in die ... → Ich sagte ihr, dass ... → so heftig, dass ich ... → fühlte sich an, als → war ich fix und fertig → Ich hatte das Gefühl, → dass es nicht mehr lange ... → ich dachte, ich muss ... → , aber es ging nicht → , was das Zeug hielt → dann ging alles ganz schnell → Ich weiß nur noch ... → um 16:38 war es → das Licht der Welt erblickte → ich konnte es nicht glauben → ich war so froh ... → , dass es vorbei war → ich hätte nie gedacht, → und ich muss sagen, → Für mich war es eine ... → noch vor sich haben ...

3 Plädoyer für mehr Experimentierfreude

Auch die Sozial- und Kulturwissenschaften tendieren dazu, Standardmethoden zu entwickeln, um bestimmte Analysen valide und reliabel durchführen zu können. Das führt z. B. zu Best Practice-Empfehlungen für die Datenanalyse und Datenvisualisierung, die Eingang in korpuslinguistische Literatur finden. Dies ist wichtig. Trotzdem: Die grundlegenden Überlegungen zu diagrammatischen Operationen und Grundfiguren, die etwa die Korpuslinguistik beherrschen, machen deutlich, wie grundlegend bestimmte methodische Zugriffe bei der Gegenstandskonstitution sind. Methoden der Visualisierung, egal ob für explorative Zwecke oder für die Präsentation, sind nicht einfach ein zusätzliches Element der Analyse, sondern prägen ganz grundsätzlich unsere Perspektive auf die Daten. Dies wurde in wissenschaftstheoretischen Arbeiten, insbesondere zu den Naturwissenschaften, verschiedentlich aufgezeigt (Knorr Cetina 2001; Böhm 2001; Rheinberger 1994; Bredekamp, Schneider & Dünkel 2008; Zittel 2011; Burri 2016). In der Linguistik sind die Reflexionen darüber, wie die im Fach herrschenden Diagrammkulturen die Gegenstandskonstitution „Sprache" prägen, noch wenig erforscht. Gerade die Korpuslinguistik muss mit ihrem datengeleiteten Zugriff eine Vorreiterrolle übernehmen, um nicht voreilig Standards der Visualisierung zu formulieren, sondern mit verschiedenen Formen zu experimentieren und diese auch wissenschaftstheoretisch zu reflektieren. Mit Lauersdorf (2018) kann man fordern: „Use all the data! View all the data! View all the combinations! View all the angles! Use all the techniques!"

Dies kann beispielsweise auch in Domänen geschehen, die auf den ersten Blick vielleicht wenig Berührungspunkte zur Korpuslinguistik haben. Ein Beispiel ist die Gesprächslinguistik oder interaktionale Linguistik. Das Gesprächstranskript spielt hier die entscheidende Rolle, um gesprochene Sprache überhaupt zu Daten zu machen, die analysiert werden können (z. B. nach GAT; Selting et al. 1998). Gespräche können jedoch auch ganz anders visualisiert werden, wie der Versuch zeigt, Gespräche mit der Figur der Jahresringe zu visualisieren (vgl. Abb. 2.14). Die grafische Figur sieht vor, die Aktanten eines Gesprächs als Positionen auf einem Kreis zu sehen. Parallel zum Ablauf des Gesprächs (die Aufnahme wird abgespielt), produziert jeder Turn an der Stelle des Aktanten einen Kreis, wobei die Länge des Turns die Größe bestimmt. Konzentrisch dazu werden die weiteren Turns des Aktanten gezeichnet, wobei die Kreise unmittelbar bei turn-Äußerung farbig gefüllt sind, mit der Zeit jedoch verblassen, aber nicht unsichtbar werden. Zusätzlich wird der Text des jeweiligen Turns kurz angezeigt.

Nach einiger Zeit wird eine Geschichte und Dynamik des Gesprächs sichtbar, wobei Abbildung 2.14 die Unterschiede dreier Gespräche augenfällig

Abb. 2.14: Visualisierung der Gesprächsdynamik mit „Jahresringen" – Gespräche 1–3.

macht: Das erste Gespräch (in Abb. 14 links) vollzieht sich hauptsächlich zwischen zwei Personen, die beide viele, relativ kurze Turns geäußert haben, daneben aber auch längere in ähnlicher Zahl. Man würde sagen, es handelt sich um eine ausgeglichene Dialogform. Beim zweiten Gespräch (in Abb. 14 in der Mitte) nehmen drei Personen am Gespräch teil, wobei der eine (links oben) das Gespräch mit sehr vielen, eher kurzen turns dominiert. In der dritten Konstellation (in Abb. 14 rechts) agieren viele Personen miteinander, wobei die eine ebenfalls das Gespräch deutlich dominiert, gefolgt von drei, vier weiteren Personen, die auch substanziell zum Gespräch beitragen.

Neue Formen der Datenvisualisierung sind nötig, da die theoretischen Überlegungen zur interaktionalen Linguistik weiter zu sein scheinen als die gängigen Formen der Gesprächstranskription. So sieht Deppermann (2014: 323) „vier Bestimmungsstücke des sprachlichen Handelns [die] unser Verständnis von ‚Pragmatik' prägen müssen: Leiblichkeit [...], Zeitlichkeit [...], Sozialität [...], Epistemizität". Zeitlichkeit meint dabei, dass sprachliches Handeln „sequenziell organisiert und simultan mit anderen Ressourcen des Handelns verknüpft" ist und deshalb „Retrospektion und Projektion [...] konstitutive Dimensionen der situierten Sinnkonstitution" sind. Anhand eines klassischen Gesprächstranskripts ist es sehr schwer, Retrospektion und Projektion abzuleiten. Der „Jahresringe"-Darstellung in Abbildung 2.14 ist aber immerhin ein Element der Retrospektion eingeschrieben: Eine Geschichte des Gesprächs ergibt sich in grafischer Form. Auch der Aspekt der Sozialität – „[s]prachliches Handeln findet in interpersonellen (Mehrpersonen-)Konstellationen statt" (Deppermann 2014: 323) – ist in der Darstellung sichtbar.

Die „Jahresringe"-Darstellung ist nur eine erste Skizze für eine die klassischen Transkriptionsformen ergänzende Datenvisualisierung, die stark ausgebaut werden müsste, um brauchbar zu sein. Ich wollte damit aber zeigen, dass neue Wege der Datenanalyse theoriegeleitet vorgehen und dabei auch neue diagrammatische Grundfiguren finden müssen, um neue Perspektiven

auf die Daten zu ermöglichen. Dabei ist auch klar, dass eine neue Visualisierung nicht den Zweck hat, die alten Fragen besser zu beantworten, sondern neue Fragen überhaupt erst ermöglicht. Der hier skizzierte „Jahresringe"-Vorschlag ist dabei zugegebenermaßen von einem deutlich korpuslinguistischen Blick auf gesprochene Sprache geprägt.

Neben methodischen Standards der Datenaufbereitung und Analyse ist es aber gerade auch in der Korpuslinguistik ein Desiderat, Experimente der Datenvisualisierung einzugehen und dabei von einer weitreichenden Definition von Diagramm auszugehen. Die Visualisierung beginnt nicht erst mit der Analyse (oder gar der Präsentation der Analyseergebnisse). Diagrammatische Überlegungen gehen bereits mit dem theoretischen Zugriff auf die Daten einher und sind ein wichtiges Element der Gegenstandskonstitution.

Literatur

Affolter, Katrin (2016): *Visualization of narrative structures*. Universität Zürich Master-Arbeit.

Barbieri, Gian Luca, Ada Cigala, Alessandro Musetti & Paolo Corsano (2012): Looking forward to the birth of a child: Tales of motherhood in forums. *International Journal of Psychoanalysis and Education IJPE* 4(2), 4–26.

Bauer, Matthias & Christoph Ernst (2010): *Diagrammatik / Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*. Bielefeld: transcript.

Bański, Piotr (2010): Why TEI stand-off annotation doesn't quite work: And why you might want to use it nevertheless. In *Balisage: The Markup Conference 2010*, Band 5, Montréal, Canada. doi: 10.4242/BalisageVol5.Banski01.

Belica, Cyril (2001ff.): *Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemischstrukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs*. Mannheim: Institut für Deutsche Sprache. http://corpora.idsmannheim.de (letzter Zugriff: 6. 11. 2017).

Bender, John B. & Michael Marrinan (2014): *Kultur des Diagramms* (Actus et imago Band VIII). Berlin: Akademie Verlag.

Bonfanti, Corrado (2012): Roberto Busa (1913–2011), pioneer of computers for the humanities. In Arthur Tatnall (Hrsg.), *Reflections on the history of computing. Preserving memories and sharing stories*, 57–61. Heidelberg: Springer.

Bostock, Michael, Vadim Ogievetsky & Jeffrey Heer (2011): D3: data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* http://vis.stanford.edu/ papers/d3 (letzter Zugriff: 6. 11. 2017).

Bredekamp, Horst (2008): Diagrammatik. In Horst Bredekamp, Birgit Schneider & Vera Dünkel (Hrsg.), *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*, 192–197. Berlin: Akademie-Verlag.

Bredekamp, Horst, Birgit Schneider & Vera Dünkel (Hrsg.) (2008): *Das Technische Bild: Kompendium zu einer Stilgeschichte wissenschaftlicher Bilder*. Berlin: Akademie Verlag.

Brezina, Vaclav, Tony McEnery & Stephen Wattam (2015): Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2), 139–173. doi: 10.1075/ijcl.20.2.01bre.
Bubenhofer, Noah (2014): Geokollokationen – Diskurse zu Orten: Visuelle Korpusanalyse. *Mitteilungen des Deutschen Germanistenverbandes* 61(1), 45–59.
Bubenhofer, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Roth, Martin Wengeler & Alexander Ziem (Hrsg.), *Handbuch Sprache in Politik und Gesellschaft* (Handbücher Sprachwissen 19), 69–93. Berlin, Boston: de Gruyter Mouton.
Bubenhofer, Noah (2018a): Serialität der Singularität. Korpusanalyse narrativer Muster in Geburtsberichten. *Zeitschrift für Literaturwissenschaft und Linguistik: Themenheft Alltagspraktiken des Erzählens*.
Bubenhofer, Noah (2018b): Diskurslinguistik und Korpora: Daten im Vektorraum. In Ingo Warnke (Hrsg.), *Handbuch Diskurs* Handbücher Sprachwissen, Berlin, Boston: de Gruyter Mouton.
Bubenhofer, Noah (2018c): Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In Noah Bubenhofer & Marc Kupietz (Hrsg.), Visualisierung sprachlicher Daten: Visual Linguistics –Praxis – Tools. Heidelberg: Heidelberg University Publishing. doi: 10.17885/heiup.345.474
Bubenhofer, Noah, Nicole Müller & Joachim Scharloth (2013): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. *Zeitschrift für Semiotik, Methoden der Diskursanalyse* 35(3–4), 419–444.
Bubenhofer, Noah & Klaus Rothenhäusler (2016): „Korporatheken“: Die digitale und verdatete Bibliothek. *027.7 Zeitschrift für Bibliothekskultur/Journal for Library Culture* 4(2), 60–71.
Bubenhofer, Noah, Klaus Rothenhäusler, Katrin Affolter & Danica Pajovic (2017): The linguistic construction of world – an example of visual analysis and methodological challenges. In Ronny Scholz (Hrsg.), *Quantifying approaches to discourse for social scientists*, Basingstoke: Palgrave Macmillan.
Burri, Regula Valérie (2016): Bilder als soziale Praxis: Grundlegungen einer Soziologie des Visuellen/Images as social practice: Outline of a sociology of the visual. *Zeitschrift für Soziologie* 37(4), 342–358. doi: 10.1515/zfsoz-2008-0404.
Busa, Roberto (1951): *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum: primo saggio di indici di parole automaticamente composti e stampi da macchine IBM a schede perforate = A 1st example of word index automatically compiled and printed by IBM punched card machines* (Archivum philosophicum Aloisianum. Serie 2). Milano: Bocca.
Böhm, Gottfried (2001): Zwischen Auge und Hand. Bilder als Instrumente der Erkenntnis. In Bettina Heintz & Jörg Huber (Hrsg.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Wellen* (Theorie:Gestaltung 01), 43–54. Wien, New York: Springer.
Chen, Chun-houh, Wolfgang Härdle & Antony Unwin (Hrsg.) (2008): *Handbook of data visualization* (Springer handbooks of computational statistics). Heidelberg: Springer.
Colloseus, Cecilia (2016): *Gebären – Erzählen. Kulturanthropologische und interdisziplinäre Perspektiven auf die Geburt als leibkörperliche Grenzerfahrung*. Mainz: Johannes Gutenberg-Universität Dissertation.
Deppermann, Arnulf (2014): Pragmatik revisited. In Ludwig Eichinger (Hrsg.), *Sprachwissenschaft im Fokus Positionsbestimmungen und Perspektiven* Jahrbuch des Instituts für Deutsche Sprache, 323–352. Berlin, Boston: De Gruyter.

Dorling, Danny (1993): Map design for census mapping. *The Cartographic Journal* 30(2), 167–183. doi: 10.1179/000870493787860175.

Echterhölter, Anna (2015): Jack Goody: Die Liste als Praktik. In Susanne Deicher & Erik Maroko (Hrsg.), *Die Liste: Ordnungen von Dingen und Menschen in Ägypten*, Band 1: Ancient Egyptian design, contemporary design history and anthropology of design, 243–261. Berlin: Kulturverlag Kadmos.

Eco, Umberto (1977): *Zeichen. Einführung in einen Begriff und seine Geschichte*. Frankfurt am Main: Suhrkamp.

Eco, Umberto (2009): *Die unendliche Liste*. München: Carl Hanser.

Ehlich, Konrad (Hrsg.) (1980): *Erzählen im Alltag* (Suhrkamp-Taschenbuch Wissenschaft 323). Frankfurt am Main: Suhrkamp.

Evert, Stefan (2009): Corpora and collocations. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics*, Band 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2), 1212–1248. Berlin, Boston: Mouton de Gruyter.

Faruqui, Manaal & Sebastian Padó (2010): Training and evaluating a german named entity recognizer with semantic generalization. In Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde & Angelika Storrer (Hrsg.), *Proceedings of KONVENS, September 6–8, 2010, Saarland University, Saarbrücken*, 129–134. Saarbrücken.

Finkel, Jenny Rose, Trond Grenager & Christopher Manning (2005): Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, 363–370. Stroudsburg, PA: Association for Computational Linguistics.

Glasze, Georg (2009): Kritische Kartographie. *Geographische Zeitschrift* 97(4), 181–191.

Graham, Shawn, Scott Weingart & Ian Milligan (2012): Getting started with topic modeling and MALLET. http://programminghistorian.org/lessons/topic-modeling-and-mallet (letzter Zugriff: 6. 11. 2017).

Gülich, Elisabeth (1980): Konventionelle Muster und kommunikative Funktionen von Alltagserzählungen. In Konrad Ehlich (Hrsg.), *Erzählen im Alltag* (Suhrkamp-Taschenbuch Wissenschaft 323), 335–384. Frankfurt am Main: Suhrkamp.

Jullien, François (2004): Die praktische Wirkkraft der Liste: von der Hand, vom Körper, vom Gedicht. In François Jullien (Hrsg.), *Die Kunst, Listen zu erstellen*, 15–50. Berlin: Merve.

Keim, Daniel A., Jörn Kohlhammer, Geoffrey Ellis & Florian Mansmann (2010): *Mastering the information age – solving problems with visual analytics*. Goslar: Eurographics Association.

Kleiner, Stefan (2011 ff.): *Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG)*. Mannheim: IDS. http://prowiki.ids-mannheim.de/bin/view/AADG/ (letzter Zugriff: 6. 11. 2017).

Knorr Cetina, Karin (2001): „Viskurse“ der Physik. Konsensbildung und visuelle Darstellung. In Bettina Heintz & Jörg Huber (Hrsg.), *Mit dem Auge denken: Strategien der Sichtbarmachung in wissenschaftlichen und virtuellen Wellen* (Theorie:Gestaltung 01), 305–320. Wien, New York: Springer.

Koplenig, Alexander (2017): A data-driven method to identify (correlated) changes in chronological corpora. *Journal of Quantitative Linguistics* 4(24), 289–318. doi: 10.1080/09296174.2017.1311447.

Kruja, Eriola, Joe Marks, Ann Blair & Richard Waters (2002): A short note on the history of graph drawing. In Petra Mutzel, Michael Jünger & Sebastian Leipert (Hrsg.), *Graph drawing* (Lecture Notes in Computer Science 2265), 272–286. Berlin, Heidelberg: Springer.

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis: Grundlinien einer Diagrammatologie*. Frankfurt am Main: Suhrkamp.

Labov, William & Joshua Waletzky (1973): Erzählanalyse. Mündliche Versionen persönlicher Erfahrung. In Jens Ihwe (Hrsg.), *Literaturwissenschaft und Linguistik*, Band 2, 78–126. Frankfurt am Main: Athenäum.

Lauersdorf, Mark Richard (2018): Linguistic visualizations as objets d'art? In Noah Bubenhofer & Marc Kupietz (Hrsg.), *Visual linguistics*, Heidelberg: Heidelberg University Publishing. [Im Druck].

Lima, Manuel (2014): *The book of trees: Visualizing branches of knowledge*. New York: Princeton Architectural Press.

Luhn, Hans Peter (1960): Key word-in-context index for technical literature (kwic index). *American Documentation* 11(4), 288–295. doi: 10.1002/asi.5090110403.

Manning, Christopher D. & Hinrich Schütze (2002): *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean (2013): Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]* http://arxiv.org/abs/1301.3781 (letzter Zugriff: 6. 11. 2017).

Naumann, Carl Ludwig (1982): Kartographische Datendarstellung. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert E. Wiegand (Hrsg.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, Band 1 (Handbücher zur Sprach- und Kommunikationswissenschaft 1), 667–692. Berlin, Boston: de Gruyter Mouton.

Peirce, Charles S. (1994): *The collected papers of Charles Sanders Peirce*. Charlottesville, VA: InteLex Corp. http://www.nlx.com/collections/95 (letzter Zugriff: 6. 11. 2017).

Perkuhn, Rainer & Cyril Belica (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. *Sprachreport* 22(1), 2–8.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. Stuttgart: UTB.

Pfeffer, Jürgen (2010): Visualisierung sozialer Netzwerke. In Christian Stegbauer (Hrsg.), *Netzwerkanalyse und Netzwerktheorie*, 227–238. Wiesbaden: Springer VS.

Pigeot, Jacqueline (2004): Die explodierte Liste: die Tradition der heterogenen Liste in der alten japanischen Literatur. In François Jullien (Hrsg.), *Die Kunst, Listen zu erstellen*, 73–121. Berlin: Merve.

Placcius, Vincentius (1689): *De arte excerpendi. Vom Gelahrten Buchhalten*. Hamburg: Gottfried Liebezeit. http://echo.mpiwg-berlin.mpg.de/MPIWG:7X92X29A (letzter Zugriff: 6. 11. 2017).

Quasthoff, Uta (1980): *Erzählen in Gesprächen: linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags* (Kommunikation und Institution 1). Tübingen: Narr.

Redder, Angelika (2001): Aufbau und Gestaltung von Transkriptionssystemen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven Sager (Hrsg.), *Text- und Gesprächslinguistik / Linguistics of text and conversation*, Band 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 16.2), 1038–1059. Berlin, Boston: de Gruyter Mouton.

Rheinberger, Hans-Jörg (1994): Experimentalsysteme, Epistemische Dinge, Experimentalkulturen Zu einer Epistemologie des Experiments. *Deutsche Zeitschrift für Philosophie* 42(3), 405–418.

Sachs, Klaus-Jürgen & Thomas Röder (1989): Partitur. In Ludwig Finscher (Hrsg.), *Die Musik in Geschichte und Gegenwart: allgemeine Enzyklopädie der Musik*, Band 10, 1424–1437.

Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson (1974): A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4), 696–735. doi: 10.2307/412243.
Sankey, Henry Riall (1896): The thermal efficiency of steam-engines. (including appendixes). *Minutes of the Proceedings of the Institution of Civil Engineers* 125(1896), 182–212. doi: 10.1680/imotp.1896.19564.
Schiller, Anne, Simone Teufel & Christine Thielen (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS. Arbeitspapier*. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Universität Tübingen, Seminar für Sprachwissenschaft.
Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (letzter Zugriff: 6. 11. 2017).
Schmid, Helmut (1995): *Improvements in part-of-speech tagging with an application to german*. Dublin. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf (letzter Zugriff: 6. 11. 2017).
Scholz, Ronny & Annika Mattissek (2014): Zwischen Exzellenz und Bildungsstreik. Lexikometrie als Methodik zur Ermittlung semantischer Makrostrukturen des Hochschulreformdiskurses. In Johannes Angermüller, Martin Nonhoff, Eva Herschinger, Felicitas Macgilchrist, Martin Reisigl, Juliette Wedl, Daniel Wrana & Alexander Ziem (Hrsg.), *Diskursforschung. Ein interdisziplinäres Handbuch*, Band 2, 86–112. Bielefeld: transcript.
Schumann, Heidrun & Wolfgang Müller (1999): *Visualisierung: Grundlagen und allgemeine Methoden*. Springer DE.
Selting, Margret, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Christoph Meier, Uta Quasthoff, Peter Schlobinski & Susanne Uhmann (1998): Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91–122.
Siegel, Steffen (2009): *Tabula: Figuren der Ordnung um 1600*. Berlin: Akademie Verlag.
Smith, Neil (1992): History and philosophy of geography: Real wars, theory wars. *Progress in Human Geography* 16(2), 257–271. doi: 10.1177/030913259201600208.
Stetter, Christian (2005): Bild, Diagramm, Schrift. In Gernot Grube, Werner Kogge & Sybille Krämer (Hrsg.), *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine* (Kulturtechnik 4), 115–136. München: Wilhelm Fink Verlag.
Steyer, Kathrin (2013): *Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht* (Studien zur Deutschen Sprache 65). Tübingen: Narr.
Zittel, Claus (2011): Ludwik Fleck und der Stilbegriff in den Naturwissenschaften. Stil als wissenschaftshistorische, epistemologische und ästhetische Kategorie. In Horst Bredekamp & John Michael Krois (Hrsg.), *Sehen und Handeln*, 171–206. Berlin, Boston: de Gruyter.

Joachim Scharloth

3 Korpuslinguistik für sozial- und kulturanalytische Fragestellungen

Grounded Theory im datengeleiteten Paradigma

Abstract: Der Beitrag entfaltet die Methodologie einer korpuslinguistischen Forschung mit kultur- und sozialwissenschaftlichen Erkenntnisinteressen. Um den Verkürzungen von Datenpositivismus und digitalem Behaviorismus zu entgehen, regt der Beitrag an, Prinzipien der Grounded Theory auf den korpuslinguistischen Forschungsprozess zu übertragen. Am Beispiel einer datengeleiteten Analyse von Bundespressekonferenzen wird illustriert, wie offenes, axiales und selektives Kodieren mit Methoden der maschinellen Textanalyse durchgeführt werden und zu einem Modell der kommunikativen Gattung führen können, das zum Verständnis ihrer gesellschaftlichen Funktion beiträgt.

Keywords: Data-driven Turn, Grounded Theory, Korpuslinguistik, Pragmatik

1 Korpuslinguistik und linguistische Kulturanalyse

Durch die Analyse von Korpora zu einem tieferen Verständnis von Kultur und Gesellschaft gelangen zu wollen, scheint ein törichtes Unterfangen. Wer Kulturen mit Clifford Geertz als „webs of significance“ (Geertz 1973: 5) betrachtet, als Menge ineinandergreifender Systeme auslegbarer Zeichen, als Texte mit komplexer Verweisstruktur, die in vergänglichen Beispielen geformten Verhaltens geschrieben sind (vgl. Geertz 1983: 15), wird schwerlich auf Textdatenbanken zurückgreifen. In Korpora sind die Äußerungskontexte der gesammelten Sprachdaten unsichtbar, sieht man von den spärlich vorhandenen Metadaten ab. Korpusanalytische Tools erlauben keinen Zugriff auf die Bedeutung, sondern lediglich auf die sprachliche Oberfläche. Sie bilden keine komplexen Verweiszusammenhänge ab, sondern liefern Listen isolierter Einzelphänomene.

Joachim Scharloth, Waseda University, School of International Liberal Studies,
1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050 Japan, E-Mail: scharloth@waseda.jp

https://doi.org/10.1515/9783110538649-004

Gängige Verfahren der Korpusanalyse nähern sich der Analyse eines relevanten sprachlichen Phänomens auch nicht, indem sie den Ko-Text einer Äußerung einbeziehen und individuell interpretieren oder kategorisieren. Ihre Ergebnisse sind aggregierte Daten oder probabilistische Modelle und die sprachwissenschaftliche Interpretation nimmt das Aggregat als Ausgangspunkt.

Gesellschafts- oder Kulturanalyse mit den Mitteln der Korpuslinguistik betreiben zu wollen, scheint also auf den ersten Blick ein fragwürdiges Unterfangen. Will man es dennoch tun, muss man sich von traditionellen Arbeitsweisen der Sozial- und Kulturanalyse verabschieden und eine neue theoretische Orientierung suchen. Die Korpuslinguistik kann keine ergänzende oder alternative Operationalisierung für Kategorien und Arbeitsweisen der traditionellen, meist qualitativ verfahrenden Kulturanalyse sein. Sie muss andere Kategorien und andere Strategien der (Re-)Konstruktion soziokultureller Bedeutungen entwickeln.

Solche Kategorien finden sich beispielsweise in der linguistischen Pragmatik: In Helmuth Feilkes Theorie der idiomatischen Prägung manifestieren sich pragmatische Informationen nicht nur auf der Ebene von Sprechakten, sondern sind zeichenhaft manifest „im pragmatischen Mehrwert oder Gebrauchswert von Einheiten aller sprachlicher Strukturbereiche" (Feilke 2000: 78). In sprachliche Muster, die als Spuren an der sprachlichen Oberfläche beobachtbar sind, hat sich demnach ein Gebrauchswert eingeschrieben.

Wie dies gemeint und begründet ist, lässt sich am besten mittels eines Blicks in die Geschichte der Sprachwissenschaft verstehen. Feilke (2003: 217 ff.) teilt die „pragmatische Wende" in der Sprachwissenschaft grob in zwei Phasen ein. Die erste Phase seit den 1970er Jahren begründete den Anspruch, sprachliche Tatbestände grundsätzlich vom Texthandeln und seinen Gelingensbedingungen her zu konzipieren und zu beschreiben (Feilke 2000: 65). In dieser Phase wird zwar der Bereich linguistischer Gegenstände und Kategorien erweitert, allerdings wurde die Pragmatik vom systemlinguistisch interessierten Zweig der Disziplin vereinnahmt: An die junge linguistische Teildisziplin wurde der Anspruch herangetragen, analog zum universalgrammatischen ein universalpragmatisches Erkenntnisinteresse zu verfolgen, das sich um den Nachweis der Universalität von Sprechakten bemühen, deren tiefenstrukturellen Gemeinsamkeiten untersuchen und sich der Erarbeitung einer allgemein gültigen Illokutionslogik verschreiben sollte (vgl. Nerlich 1995: 311). In der zweiten Phase der pragmatischen Wende, die die letzten 25 bis 30 Jahre linguistischer Forschung prägte, wurden unterschiedliche pragmatische Positionen neu formuliert (Feilke 2003: 217 ff.). Für die Korpuspragmatik sind dabei zwei Aspekte besonders relevant: (1) Zum einen rückte die Formelhaftigkeit des Sprechens und Schreibens, die über die sprachliche Oberfläche zugänglich ist, zulasten

sprachlicher Universalien der Tiefenstruktur ins Zentrum der Theoriebildung. (2) Zum anderen wird der Kontext einer sprachlichen Äußerungen nicht mehr als quasi-objektiv gegebene Interpretationsressource betrachtet, mit deren Hilfe sich die semantische Unterbestimmtheit und Mehrdeutigkeit einer Äußerung konkretisieren ließen. Vielmehr geht die neuere Pragmatik davon aus, dass der Sprachgebrauch den Kontext (mit-)herstellt („kontextualisiert", vgl. Bubenhofer 2009), dass sprachliche Routineformeln mithin als Kontextualisierungshinweise gelesen werden können.

Idiomatische Prägungen können mit Feilke (2003) als Resultat von Konventionalisierungen von Interpretationen verstanden werden, die diese Konventionalisierung ausdrucksseitig widerspiegeln. Es sind demnach Indikatoren auf der sprachlichen Oberfläche, die einen Text als Sprachhandlung situieren – etwa „hier ist es schön" in Kombination mit „mir geht es gut", die sofort auf eine Ansichtskarte schließen lassen. Signifikant häufig auftretende sprachliche Muster können deshalb als das Ergebnis rekurrenter Sprachhandlungen der Sprecherinnen und Sprecher gedeutet werden, in die typische Verwendungskontexte, Handlungsziele und Interpretationsrahmen eingeschrieben sind.

2 Der Data-driven Turn

Für die wissenschaftliche Identifizierung rekurrenter sprachlicher Formen sind korpuslinguistische Methoden hochgradig geeignet. Rekurrenz, also das wiederholte Auftreten sprachlicher Formen in Texten, ist als Frequenz sprachlicher Einheiten in Korpora operationalisierbar. Und rekurrente Muster können induktiv ermittelt werden. Während die Korpuslinguistik in der systemorientierten Linguistik die Funktion hat, wiederkehrende Muster des Sprachgebrauchs zu identifizieren, die dann als Regularitäten oder Gebrauchsnormen gedeutet werden, werden in den kultur- und sozialwissenschaftlich interessierten Zweigen der Linguistik rekurrente sprachliche Muster mit kulturellen oder sozialen Phänomenen in Zusammenhang gebracht, je nach sprachtheoretischer Haltung werden sie entweder als deren Symptom oder als diese (mit-)konstituierend gedeutet.

Bereits 1957 rückte Firth (1957: 194) mit dem Begriff „collocation" ein Phänomen in das Zentrum linguistischer Reflexion, das eigentlich arbiträre Wortverbindungen zu usuellen umdeutete. Kollokationen lassen sich nicht über grammatische Regeln und semantisches Wissen über Einzellexeme erschließen, sondern müssen gelernt werden. In Kollokationen, verstanden als überzufällig häufig auftretende Kookkurenz sprachlicher Einheiten, die sich durch statistische Analysen der sprachlichen Oberfläche leicht identifizieren lassen

(Evert 2009), ist in dieser Perspektive also Sprachgebrauchswissen eingeschrieben. Das Interesse für Kollokationen entfaltete sich in der Folge einerseits im Kontext der Phraseologie (Burger 1998), andererseits in der Computerlinguistik und verwandten Gebieten (Carstensen et al. 2010). Während die phraseologische Perspektive Anwendungsbezüge im Bereich der Vermittlung von Fremdsprachen entwickelte, in denen Sprachgebrauchswissen und nicht (nur) Wissen über Sprachsysteme vermittelt werden sollten (vgl. Hausmann 1985), nutzte die Computerlinguistik Kollokationsanalysen für datengeleitete Modelle in der Semantik und für die Entwicklung von Methoden der Textklassifikation. Die Kategorie der Kollokation steht exemplarisch für eine Reihe analytischer Kategorien wie Keyword oder n-Gramm, die sich im Zuge der Verbreitung korpuslinguistischer Methoden in andere Felder der Linguistik als Standardanalysekategorien etabliert und die allesamt gemeinsam haben, dass sie leicht über die sprachliche Oberfläche identifizierbar sind und dass ihnen ein pragmatischer Mehrwert zugeschrieben wird.

Als Katalysator der Durchsetzung maschineller Methoden der Text- und Sprachanalyse und damit auch der Korpuslinguistik hat sich die voranschreitende digitale Revolution erwiesen. Sie verändert auch die Geistes-, Kultur- und Sozialwissenschaften und insbesondere deren Umgang mit Texten. Unter dem Begriff der digitalen Revolution können im Hinblick auf die Kultur- und Sozialwissenschaften das Zusammenwirken dreier tiefgreifender Wandelprozesse verstanden werden:

1. Die Verdatung der Welt: Immer mehr Informationen entstehen in digitalen Formaten, ehemals analoge Daten werden digitalisiert. „Digital" bedeutet dabei zunächst „abzählbar sein", d. h. dass Informationen in eine numerische Form gebracht und mit mathematischen Methoden analysierbar werden. Parallel zur Entstehung von *Big Data* ermöglicht die Digitalisierung damit auch
2. die Zusammenführung und damit kombinierte Analyse von Daten unterschiedlichster Provenienz: die Repräsentation unterschiedlichster Informationstypen in einem numerischen Modell macht es möglich, unterschiedlichste Informationen durch Algorithmen miteinander zu verknüpfen und zu analysieren. Dies ist die Grundlage für
3. die zunehmende Emanzipation der Daten von dem Zweck ihrer Produktion: war bislang der Aufbau eines Datenarchivs eng mit einem Zweck verknüpft, der in der Struktur des Archivs und seiner Findemittel sichtbar wurde, erlaubt die Digitalisierung nun jede in einem mathematischen Modell mögliche Anfrage an die Daten und damit die Emanzipation des Nutzers von den Strukturen des Archivs. Damit verbunden ist allerdings auch ein Kontrollverlust im Sinne einer Verfügungsmacht über die eigenen Daten (vgl. Scharloth, Eugster & Bubenhofer 2013).

Korpora können als Nachschlagewerke benutzt werden, um zu überprüfen, ob sich darin Belege für ein bestimmtes sprachliches Phänomen finden. So kann beispielsweise geprüft werden, ob ein sprachliches Phänomen, von dem vermutet wird, dass es typisch für einen bestimmten Kontext ist, in einem entsprechenden Kontext auch signifikant häufiger vorkommt als in anderen Kontexten. Korpora haben dann die gleiche Funktion wie Zettelkästen früher in der Lexikographie: Sie liefern Belegstellen, die Linguistinnen und Linguisten abhängig von ihrem Erkenntnisinteresse interpretieren. So wenig diese Art, Forschung zu betreiben, für sich genommen problematisch ist, so wenig nutzt sie aber die durch die Digitalisierung entstandenen neuen Möglichkeiten. Sie ist im Prinzip eine Fortsetzung der bisherigen Forschung mit effizienteren, digitalen Mitteln, weswegen sie im anglophonen Kontexten auch als „digitized humanities" (im Gegensatz zu „digital humanities") bezeichnet wird.

Doch schon 2006 fragten Perkuhn und Belica: „Sind Korpora nur Belegsammlungen oder Zettelkästen in elektronischer Form?" (Perkuhn & Belica 2006: 2) Und ihre Antwort fiel angesichts wachsender Kapazitäten und Prozessorgeschwindigkeiten sowie der wachsenden Verfügbarkeit elektronischer Korpora deutlich aus: „Mitnichten! In entsprechender Größe [...] und mit den entsprechenden Analysemethoden eröffnen sie eine eigene Perspektive in der linguistischen Forschung – die korpuslinguistische Perspektive." (Perkuhn & Belica 2006: 2) Diese Perspektive überprüft nicht theoretisch begründete Hypothesen mittels Korpusdaten, sondern sucht induktiv nach Mustern in großen Sprachdatenmengen, um so zu neuen Einsichten über Sprache zu gelangen und neue Beschreibungskategorien zu entwickeln. Wolfgang Teubert formuliert das erkenntnistheoretische Programm des „data-driven turn" (Scharloth, Eugster & Bubenhofer 2013) in der Linguistik wie folgt:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach. (Teubert 2005: 4)

Dabei entspricht der induktive Zugang einer Hinwendung zu einer genuin digitalen Forschungslogik.

Die induktive Vorgehensweise ist auch auf Fragen aus dem Bereich der linguistischen Pragmatik anwendbar. Korpuspragmatisch zu forschen bedeutet demnach, in großen Textkorpora induktiv nach signifikant häufig auftretenden Mustern zu suchen und diese Muster als Ausdruck von rekurrenten Sprachhandlungen der Autorinnen und Autoren der im Korpus enthaltenen Texte bzw. der sie autorisierenden Institutionen und Gruppen zu interpretieren, mithin als Muster mit soziokultureller Salienz.

3 Der Datenbegriff der Korpuslinguistik

Datengeleitete, strukturentdeckende Verfahren erfreuen sich in Zeiten von „Künstlicher Intelligenz“ und „Deep Learning“ großer Beliebtheit. Mit ihnen ist die Hoffnung verknüpft, Theoriebildung durch Computermodelle zu ersetzen, deren Tauglichkeit sich daran messen lassen muss, ob mit ihnen die gewünschten lebensweltlichen Zwecke erfüllt werden können oder nicht. Für Sozial- und Kulturwissenschaftler, die einen eher verstehenden Zugang zur Welt haben, mag der rein algorithmische Zugriff auf ganze Lebensbereiche befremdlich wirken, er ist aber gleichwohl prägend für einen wesentlichen Teil der IT-Branche. Als repräsentativ für diese neue Spielart des Positivismus kann ein programmatischer Essay von Chris Anderson, dem Chefredakteur des *Wired Magazine*, gelten. In seinem „The End of Theory“ betitelten Aufsatz (Anderson 2008) verkündete er angesichts von Big Data (der reinigenden „Datensintflut“) das bevorstehende Ende traditioneller wissenschaftlicher Erkenntnisprozesse. Das Anhäufen großer Datenschätze und die Möglichkeit ihrer effizienten Analyse mache wissenschaftliche Methoden überflüssig, in seinen Worten: „[F]aced with massive data, this approach to science – hypothesize, model, test – is becoming obsolete.“ (Anderson 2008: o. S.) Das empirisch-analytische Wissenschaftsbild ist nach Meinung Andersons obsolet geworden, weil die Zahlen für sich selbst sprechen:

> Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson 2008: o. S.)

Für die Korpuslinguistik würde dies bedeuten: Linguistische Theorie wird ersetzt durch hochgradig ausdifferenzierte wahrscheinlichkeitsbasierte Kalküle, die als Ergebnis der Analyse großer Datenmengen anfallen. Grammatiktheorien bedürfen keiner Kategorien wie Subjekt, thematischer Rolle oder Modalverb mehr; solange das algorithmisch aus einem Trainingsdatensatz generierte Modell wohlgeformt erscheinende Sätze zu produzieren in der Lage ist, müssen wir dieses Modell auch nicht verstehen. Sprachliches Handeln ist demnach berechenbar, wenn für das System nur genügend Trainingsdaten zur Verfügung stehen.

Andersons Utopie einer theoriefreien Welt beruht auf den Implikationen der Wendung „[w]ith enough data, the numbers speak for themselves“ (Anderson 2008: o. S.), die in mehrfacher Hinsicht problematisch ist. Sie birgt einerseits das Potenzial, kritikimmunisierende ad-hoc Modifikationen (Popper 1966: 57 f.) zu generieren: Wenn ein stochastisches Simulationsmodell nicht funktioniert,

könnte immer der Einwand erhoben werden, es hätten nicht genügend Daten vorgelegen. Andererseits verdankt die Wendung „[w]ith enough data, the numbers speak for themselves“ ihre vermeintliche Plausibilität einem verkürzten „data“-Begriff: Daten sind durch Messung oder Beobachtung zustande gekommene Repräsentationen, die in irgendeiner Form strukturiert sein müssen, um mit Hilfe von Codes als Informationen lesbar zu sein. Keine Beobachtung jedoch kommt ohne Theorie aus, keine Strukturierung ohne begrifflich-theoretische Kategorisierung: „Es gibt keine reinen Daten.“ (Stachowiak 1973: 288) In die Bestimmung dessen, was überhaupt gemessen werden soll, fließt also theorieförmiges Wissen ein. Nur wenn überhaupt *relevante* Merkmale erfasst werden, können die Ergebnisse datengeleiteter Analysen valide sein.

Befindet sich die Korpuslinguistik aber auf demselben positivistischen Holzweg wie der Herausgeber des *Wired Magazine*, wenn Wolfgang Teubert fordert, dass es „the discourse itself“ sein müsse, „which will have to provide the categories and classifications that are needed to answer a given research question“ (Teubert 2005: 4)? Will die Korpuslinguistik den Eigenheiten ihres Gegenstandes gerecht werden und will sie Antworten auf kultur- und sozialwissenschaftliche Fragen geben können, dann liegt auf der Hand, dass sie ihren Datenbegriff grundlegend anders orientieren und vor allem einen anderen analytischen Zugang zu diesen Daten wählen muss.

Für korpuslinguistische Daten gilt wie für alle Daten in den Kultur- und Sozialwissenschaften, dass es sich um interpretierte Daten handelt und dass Aussagen über diese Daten Interpretationen von Interpretationen sind. Clifford Geertz stellte in seinem Essay „Thick Description“ fest:

> In finished anthropological writings, including those collected here, this fact – that what we call our data are really our own constructions of other people's constructions of what they and their compatriots are up to – is obscured because most of what we need to comprehend a particular event, ritual, custom, idea, or whatever is insinuated as background information before the thing itself is directly examined. [...] There is nothing particularly wrong with this, and it is in any case inevitable. [...] Right down at the factual base, the hard rock, insofar as there is any, of the whole enterprise, we are already explicating: and worse, explicating explications. (Geertz 1973: 9)

Kulturwissenschaftliche (und damit auch sprachwissenschaftliche) Daten sind also schon in ihrer Entstehung nicht eindeutig oder objektiv gegeben, sondern immer schon zeichenhaft. Ihre Interpretation muss daher in Form dichter Beschreibung erfolgen, die versucht, den Sinn zu erfassen, den die Handelnden diesen Daten selbst zugeschrieben haben. Maschinelle, insbesondere datengeleitete Verfahren resultieren aber fast immer in „dünnen Beschreibungen“.

Doch wie kann der Anspruch eingeholt werden, Korpusdaten als interpretierte Daten aufzufassen und als Ausdruck kultureller Relevanzen zu deuten? Und welches Verhältnis hat eine datengeleitete Korpuspragmatik zur Theorie?

4 Grounded Theory

Ich möchte vorschlagen, die Verkürzungen, die der Datenpositivismus mit sich bringt, durch eine Orientierung an Prinzipien der *Grounded Theory* zu vermeiden. Die Grounded Theory ist eine Methodologie, die in den 1960er Jahren von Barney Glaser und Anselm Strauss entwickelt wurde und zwar aus dem Unbehagen heraus, dass in der sozialwissenschaftlichen Forschungslogik dem empirischen Testen von Theorien die größte Bedeutung beigemessen wurde. In ihrem Buch *The Discovery of Grounded Theory* (1967) fordern sie dagegen, die Generierung von Theorien stärker ins Zentrum sozialwissenschaftlichen Forschens zu rücken und zwar „the discovery of theory from data" (Glaser & Strauss 1967: 1). Analog zur hier vorgestellten Spielart der Korpuslinguistik plädieren auch sie ein für induktives, datengeleitetes Vorgehen.

Grounded Theory wird heute zumeist als Methode der qualitativen Sozialforschung wahrgenommen. Das sahen Glaser und Strauss jedoch durchaus anders. Ihrer Meinung nach gibt es

> [...] no fundamental clash between the purposes and capacities of qualitative and quantitative methods or data. What clash there is concerns the primacy of emphasis on verification or generation of theory [...] We believe that each form of data is useful for both verification and generation of theory [...] In many instances, both forms of data are necessary." (Glaser & Strauss 1967: 17–18).

Die Prinzipien der Grounded Theory sind daher prinzipiell auch auf eine kultur- und sozialwissenschaftlich interessierte Korpuslinguistik als quantitativem Verfahren übertragbar. Die Generierung von Theorien aus der Auseinandersetzung mit Daten erfolgt in der Grounded Theory in einem dreistufigen Prozess des Kodierens. Unter Kodieren wird die Einordnung von Phänomenen oder Ereignissen in ein kategorial-theoretisches Vokabular verstanden, häufig eine Taxonomie (vgl. Breuer 2010: 69 ff.).

Bei der ersten Stufe, dem *offenen Kodieren*, handelt es sich um einen „analytischen Prozeß [...], durch den Konzepte identifiziert und in Bezug auf ihre Eigenschaften und Dimensionen entwickelt werden. [...] Ähnliche Ereignisse und Vorfälle werden benannt und zu Kategorien gruppiert." (Strauss & Corbin 1996: 54 f.) In der Auseinandersetzung mit den Daten sollen also zunächst relevante Phänomene identifiziert und begrifflich gefasst werden, wobei die Forschenden grundsätzlich offen für unterschiedliche Lesarten sein müssen.

Beim *axialen Kodieren*, der zweiten Stufe des Kodierprozesses, liegt der Fokus darauf, die auf der ersten Stufe identifizierten Kategorien „in Bezug auf die Bedingungen zu spezifizieren, die das Phänomen verursachen; den Kontext [...], in den das Phänomen eingebettet ist; die Handlungs- und interaktionalen

Strategien, durch die es bewältigt, mit ihnen umgegangen oder durch die es ausgeführt wird; und die Konsequenzen dieser Strategien“ (Strauss & Corbin 1996: 76). Durch den systematischen Vergleich der Kontexte und pragmatischen Dimensionen der identifizierten Kategorien wird es möglich, diese zu gruppieren und ggf. zu hierarchisieren.

Das *selektive Kodieren* als dritte und letzte Stufe des Kodierprozesses ist letztlich wieder ein axiales Kodieren, allerdings auf einer höheren Abstraktionsstufe. Ziel des selektiven Kodierens ist es, eine Schlüsselkategorie zu identifizieren, die als eine Art konzeptuelles Zentrum der zu entwickelnden Theorie fungiert (vgl. Strübing 2008: 20 f.). Im Anschluss werden die übrigen Kategorien um das Zentralkonzept herum gruppiert und auf die Qualität der Verknüpfung mit diesem hin untersucht (vgl. Breuer 2010: 92). Das so gewonnene Modell kann als eine bereichsbezogene Theorie gelten.

Die Darstellung des Kodierungsprozesses erfolgte hier zwar stufenweise, allerdings darf und soll der Prozess des Kodierens nicht ausschließlich sukzessiv erfolgen, vielmehr entfaltet er sein Potenzial erst im Rahmen der „ausgebauten konsekutiv-iterativ-rekursiven Strategie des Hin und Her, des Vor und Zurück zwischen Datenerhebung, Konzeptbildung, Modellentwurf und Modellprüfung sowie der Reflexion des Erkenntniswegs“ (Breuer 2010: 69).

Wie sich dieser zugegebenermaßen abstrakte Prozess in der korpuslinguistischen Praxis konkretisieren kann, soll im Folgenden anhand einer datengeleiteten Analyse der kommunikativen Gattung der Regierungspressekonferenz illustriert werden.

5 Das Korpus

Die kommunikative Gattung der Regierungspressekonferenz ist gesprächsstrukturell durch Frage-Antwort-Sequenzen geprägt (Klein 2001: 1600). Diese werden durch die Gesprächsrollen Politiker/in (und/oder ihre Sprecher/in) und Journalist/in andererseits gefüllt. Regierungspressekonferenzen sind in den meisten Fällen nicht monothematisch, sondern behandeln sukzessiv mehrere verschiedene Themen. Regierungspressekonferenzen werden in Deutschland nicht von der Regierung organisiert, sondern vom Verein Bundespressekonferenz (BPK). Dabei handelt es sich um einen unabhängigen Zusammenschluss der deutschen Parlamentsjournalistinnen und -journalisten. Dieser Verein lädt üblicherweise dreimal wöchentlich Vertreterinnen und Vertreter der Bundesregierung zu Pressekonferenzen ein (vgl. Bundesregierung 2016). Aus der Perspektive der Public-Relations-Theorie sind Regierungspressekonferenzen Regierungs-PR. Als solche wird ihnen einerseits eine aufklärungs- und infor-

Tab. 3.1: Überblick über das Regierungspressekonferenzen-Korpus.

Jahr	Anzahl Pressekonferenzen	Anzahl Token
2013	12	49.559
2014	140	836.278
2015	142	113.572
2016	94	720.040

mationsorientierte Funktion, andererseits aber auch eine herrschafts- und machtorientierte Funktion im Sinne der Herstellung von Akzeptanz, Vertrauen und Verständnis für politische Maßnahmen und Ziele zugeschrieben (vgl. Köhler & Schuster 2006: 18 f.).

Für die folgende Analyse wurden die Wortlautprotokolle der Regierungspressekonferenzen von der Website der Bundesregierung (www.bundesregierung.de [letzter Zugriff: 15. 6. 2017]) heruntergeladen und so geparst, dass der Wechsel von Turns und die turnspezifische Sprecherrolle (Regierungsvertreter/in oder Journalist/in) annotiert werden konnten. Insgesamt enthält das Korpus 388 Regierungspressekonferenzen mit insgesamt 2.721.862 laufenden Wortformen vom Dezember 2013 bis August 2016 (vgl. Tab. 3.1).

Von den insgesamt 41.988 Turns entfielen auf Journalistinnen und Journalisten 21.721 mit insgesamt 975.824 Wortformen, auf die Regierungsvertreterinnen und -vertreter 20.267 Turns mit zusammen 1.746.038 Wortformen. Dabei waren die Fragen erwartungsgemäß im Durchschnitt deutlich kürzer als die Antworten: Turns von Regierungsvertreterinnen und vertretern waren mit im Durchschnitt 86.2 Wortformen fast doppelt so lang wie die der Pressevertreter/innen mit 44.9 Wortformen.

Das Korpus wurde mit Hilfe des TreeTaggers (Schmid 1994) tokenisiert, mit Wortarten-Informationen annotiert und lemmatisiert. Beim verwendeten Tagset handelt es sich um das Stuttgart-Tübingen-Tagset (STTS) (Schiller, Teufel & Thielen 1995). Darüber hinaus wurden Satzgrenzen und Nominalgruppen mithilfe eigener Scripts identifiziert und annotiert.

6 Exemplarische Analyse

Das Kodieren, das die Praxis der Grounded Theory prägt, übersetzt sich in der datengeleiteten Korpuslinguistik in Prozesse der Identifizierung signifikanter Muster (im Folgenden „mustern“ genannt), ihrer typischen Kontexte und schließlich in der Synthese zu einem Modell. Dieses Modell soll die konstitutiven Ele-

mente des Untersuchungsgegenstandes sichtbar machen und damit eine Interpretation seiner soziokulturellen Bedeutung nahelegen.

6.1 Offenes Kodieren: Standardverfahren zur Musteridentifikation

Der erste Zugriff auf die Musterhaftigkeit eines Korpus erfolgt mittels korpusanalytischer Standardverfahren, die erste Hinweise auf potenziell interessante sprachliche Merkmale im Korpus geben können. Solche Standardverfahren zum Mustern eines Korpus sind beispielsweise Keyword- oder n-Gramm-Analysen (vgl. Bubenhofer 2017; Bubenhofer & Scharloth 2016). Für eine erste Annäherung an das Untersuchungskorpus habe ich den Wortschatz in den Regierungspressekonferenzen mit dem Wortschatz in Reden Angela Merkels verglichen (vgl. Abb. 3.1). Auch wenn es sich bei Politikerreden um keine dialogische Gattung handelt, sind doch sowohl inhaltliche Aspekte als auch die mediale Mündlichkeit vielversprechende Bedingungen dafür, dass bei einem Vergleich die lexikalischen Besonderheiten der Regierungspressekonferenzen sichtbar werden, während die Gemeinsamkeiten mit anderen Gattungen der Regierungs-PR als nicht-signifikant ausgeblendet bleiben.

Hinweis Kenntnis **Interview** Stellung Zeitpunkt Überlegung Behörde Sprecherin klar **hinzufügen** Entscheidung
Finanzministerium militärisch Maßnahme nachreichen **ansprechen** BMI Innenministerium **Vorschlag**
ausführen Kanzleramt Wirtschaftsministerium verweisen aktuell betreffen **beziehen**
Reise **berichten** **Aussage** **ankündigen** mehrfach Plan entsprechend **ergänzen**
informieren genau Einschätzung zuständig Kontakt vorliegen
Gesetzentwurf Detail kurz Kabinett Freitag Mittwoch Position
bekannt Erkenntnis Person Sprecher Sicht **bewerten** Stand
Vereinbarung angehen Moment planen tatsächlich Frau
Information **kommentieren** **beantworten** Lage Fall
Wochenende prüfen **Bericht** Zahl Pressekonferenz
Haltung Ministerium **Äußerung** Termin **bestätigen**
Thema grundsätzlich Uhr Regierung Treffen
auswärtig Amt **äußern** **sagen** Herr Woche
konkret Kanzlerin Minister Außenminister
Bundeskanzlerin **Gespräch** Bundesregierung
Frage geben

Abb. 3.1: Nomen, Adjektive und Verben mit dem höchsten Signifikanz-Wert beim Vergleich des Regierungspressekonferenz-Korpus mit den Reden Angela Merkels. Lexeme, die auf Sprechhandlungen verweisen, sind fett gedruckt.

Tab. 3.2: Häufigste Tetragramme in den Turns von Journalist/innen und Regierungsvertreter/innen.

<table>
<tr><th>Journalist/innen</th><th>Regierungsvertreter/innen</th></tr>
<tr><td>Ich Sie richtig verstanden (58)</td><td>kann ich Ihnen nicht (378)</td></tr>
<tr><td>Herr Seibert können Sie (49)</td><td>Ich kann Ihnen sagen (231)</td></tr>
<tr><td>Können Sie sagen ob (49)</td><td>ich Ihnen nicht sagen (189)</td></tr>
<tr><td>Ich habe (eine | eine wichtige) Frage an Herrn Seibert (47)</td><td>kann ich Ihnen sagen (162)</td></tr>
<tr><td>Können Sie sagen wie (46)</td><td>Ich weiß nicht ob (148)</td></tr>
<tr><td>Ich weiß nicht ob (43)</td><td>kann Ihnen sagen dass (145)</td></tr>
<tr><td>ich es richtig verstanden (37)</td><td>Ich kann Ihnen nicht (129)</td></tr>
<tr><td>wenn ich es richtig (36)</td><td>kann ich nicht sagen (89)</td></tr>
<tr><td>Herr Seibert ich habe (eine | eine allgemeine) Frage (35)</td><td>Sie wissen dass es (78)</td></tr>
<tr><td>Wenn ich es richtig (32)</td><td>kann Ihnen nicht sagen (76)</td></tr>
<tr><td>Herr Seibert Sie gesagt (30)</td><td>ich Ihnen sagen kann (76)</td></tr>
<tr><td>Sie richtig verstanden dass (30)</td><td>Sie wissen dass wir (72)</td></tr>
<tr><td>Herr Schäfer können Sie (30)</td><td>kann ich sagen dass (71)</td></tr>
<tr><td>Können Sie mir sagen (29)</td><td>Wir gehen aus dass (69)</td></tr>
<tr><td>Herr Seibert gibt es (29)</td><td>Ich kann sagen dass (67)</td></tr>
<tr><td>Herr Seibert Sie sagten (29)</td><td>Sie können ausgehen dass (63)</td></tr>
<tr><td>Können Sie bestätigen dass (29)</td><td>weiß nicht ob Sie (62)</td></tr>
<tr><td>können Sie sagen wie (28)</td><td>Ich glaube nicht dass (57)</td></tr>
<tr><td>Ich wollte fragen ob (28)</td><td>Ich gehe aus dass (57)</td></tr>
<tr><td>Können Sie sagen wann (27)</td><td>ich Ihnen nicht nennen (55)</td></tr>
<tr><td>Ich wollte wissen ob (26)</td><td>ich Ihnen sagen dass (54)</td></tr>
<tr><td>Sie gesagt dass Sie (23)</td><td>Sie können sicher dass (49)</td></tr>
<tr><td>ich es richtig dass (23)</td><td>kann ich nicht bestätigen (46)</td></tr>
<tr><td>ich weiß nicht ob (23)</td><td>was ich Ihnen sagen (43)</td></tr>
<tr><td>wenn ich richtig verstanden (22)</td><td>ich ehrlich gesagt nicht (41)</td></tr>
<tr><td>Wenn Sie sagen dass (21)</td><td>Ich Ihnen gesagt dass (41)</td></tr>
<tr><td>Herr Seibert könnten Sie (21)</td><td>Ich bitte (um | um Ihr) Verständnis dass (40)</td></tr>
<tr><td>Sie sagen dass Sie (21)</td><td>kann ich nicht kommentieren (37)</td></tr>
<tr><td>Herrn Seibert Herrn Schäfer (21)</td><td>bitte (um | um Ihr) Verständnis dass ich (37)</td></tr>
<tr><td>Können Sie kurz sagen (20)</td><td>ich weiß nicht ob (36)</td></tr>
</table>

Obwohl auch in Politikerreden Kommunikationsverben ein hochsignifikantes Merkmal sind, zeigt die Analyse des typischen Vokabulars, dass Kommunikationsverben und überhaupt Lexeme, die auf Sprechhandlungen verweisen, in der nach Signifikanz geordneten Visualisierung überaus zahlreich vertreten sind. Allerdings sind es in der Regierungspressekonferenz vorwiegend Kommunikationsverben, die auf Prozesse der Informationsvergabe referieren („bestä-

tigen“, „beantworten“, „informieren“, „ankündigen“, „ergänzen“, „hinzufügen“, „ausführen“) und nicht solche, die eine Einstellung oder eine Emotion zum Ausdruck bringen („freuen“, „ärgern“, „glauben“, „meinen“), wie sie häufig in Politikerreden verwendet werden.

Eine weitere mögliche erste Annäherung an das Untersuchungskorpus ist eine n-Gramm-Analyse. N-Gramme sind Einheiten einer durch n bestimmten Anzahl aufeinanderfolgender Wörter (Manning & Schütze 2002: 192 ff.). Normalerweise werden n-Gramme als kontinuierliche Wortfolgen verstanden; um die Variation etwas zu verringern, haben ich Nominalphrasen auf ihren Kopf reduziert. Die Listen der frequentesten Tetragramme (4-Wort-Einheiten) für die Turns von Journalistinnen und Journalisten einerseits und die von Regierungsvertretern andererseits finden sich in Tabelle 3.2.

Bei den Redeanteilen der Journalist/innen dominieren n-Gramme, die auf Praktiken der Verständnissicherung, der Bitte um Auskunft und des Ausdrucks von Nichtwissen verweisen. Dagegen sind bei den Regierungsvertreter/innen n-Gramme am häufigsten, die die (Un-)Fähigkeit zur Vergabe einer Information oder die (Un-)Möglichkeit im Sinne einer (fehlenden) Autorisierung thematisieren. Dabei ist die hohe Frequenz von n-Grammen, die auf eine explizite Nicht-Vergabe von Informationen verweisen bzw. diese rechtfertigen für eine informationsorientierte kommunikative Gattung überraschend hoch.

Das Mustern des Korpus mittels induktiver Standardverfahren als Praxis des offenen Kodierens zeigt damit ein Spannungsverhältnis zwischen Mustern, die auf Sprechakte der Informationsnachfrage und -vergabe verweisen, und solchen Mustern, die die explizite Verweigerung von Information ausdrücken. Die generelle Dominanz dieser Muster verweist aber darauf, dass epistemische Aspekte prägend für die untersuchte Gattung sind. Datengeleitet zu arbeiten bedeutet nun, diese Muster als soziokulturell bedeutsame Merkmale des untersuchten Korpus aufzufassen und im nächsten Schritt des axialen Kodierens einer genaueren kontextsensitiven Analyse zu unterziehen.

6.2 Axiales Kodieren: Musterkontexte analysieren

Kontexte werden korpuslinguistisch häufig mit Hilfe von Kollokationsanalysen erschlossen. Mit „Kollokationen“ sind üblicherweise binäre Verbindungen von Wortformen oder Lemmata gemeint, die innerhalb eines „Fenster“ von x Wörtern typischerweise miteinander vorkommen (Evert 2009). Im datengeleiteten Paradigma berechnet man Kollokationen nicht isoliert, sondern grundsätzlich die Kollokationen aller Einheiten zu allen Einheiten (vgl. Scharloth, Eugster & Bubenhofer 2013). Da das untersuchte Korpus ein dialogisches und damit die Sequenzialität der Äußerungen ein konstitutives Merkmal für die sprachliche

Tab. 3.3: Nach Frequenz geordnete Muster in der Abfolge von Kommunikationsverben im Journalisten-Turn und Kommunikationsverben im darauffolgenden Turn von Regierungsvertretern.

Adjazenzmuster	Frequenz
sagen können → nicht sagen können	89
sagen → nicht sagen können	41
sagen können → sagen können	40
sagen → sagen	36
nicht sagen können → nicht sagen können	36
nicht sagen → nicht sagen können	36
sagen → sagen können	32
sagen können → nicht sagen	23
sagen können → sagen	22
sprechen → nicht sagen können	22
sagen → nicht sagen	21
nicht sagen können → sagen können	19
nicht sagen müssen → nicht sagen können	18
fragen → sagen können	16
nicht sagen → sagen	16
sagen sollen → nicht sagen können	16
nicht sagen → sagen können	15
nicht sagen wollen → nicht sagen können	15
nicht sagen können → sagen	13
sagen → nicht sagen müssen	13

Gestalt ist, erscheint es sinnvoll, Adjazenzmuster zu untersuchen, d. h. Kollokationen, in denen die kookkurierenden Einheiten in aufeinanderfolgenden Turns vorkommen.

Als potenziell kookkurrierende Einheiten wurden Kommunikationsverben zusammen mit sie begleitenden Negationswörtern und Modalverben (bspw. „nicht sagen können") jeweils zusammengefasst. Um die Kommunikationsverben in den Regierungspressekonferenzen möglichst umfassend zu analysieren, wurden sämtliche im *Handbuch deutscher Kommunikationsverben* (Harras et al. 2004) verzeichneten in die Kollokationsanalyse einbezogen

Das frequenteste Adjazenzmuster ist eine Journalistenfrage mit „sagen können" als verbalem Bestandteil in der Journalistenfrage und „nicht sagen können" in der Antwort durch den Regierungsvertreter (Tab. 3.3).

In den zehn frequentesten Adjazenzmustern finden sich fünf, in denen die explizite Unmöglichkeit bzw. Nicht-Autorisiertheit zur Vergabe von Informationen thematisiert wird. Wenn man die Adjazenzen der frequentesten Kommunikationsverben in den Frageturns untersucht, erhält man ein ähnli-

ches, sogar noch klareres Bild. Auf Turns von Journalisten, die „sagen können" (872-mal) oder „nicht sagen können" (749-mal) enthalten, dominieren in den folgenden Turns von Regierungsvertretern Formulierungen mit „nicht sagen können", „nicht sagen", „nicht sagen müssen", „nicht nennen können", „nicht beantworten können", „nicht sprechen können", „nicht sagen wollen", „nicht sagen mögen", „nicht informieren können" und „nicht sagen sollen".

Dass die Ergebnisse durchaus verallgemeinerbar für die Gattung der Regierungspressekonferenz sind, zeigt sich, wenn man sämtliche Adjazenzen auswertet. Von den 16.782 Turnabfolgen, die dem Muster Journalist/in → Regierungsvertreter/in entsprechen, konnten in 9.949 Fällen in beiden Turns Kommunikationsverben identifiziert werden, das sind 59,3 % aller Fälle. Die Analyse deckt also einen recht großen Teil des Gesamtmaterials ab. Von diesen enthielt die Antwort des Regierungssprechers in 6.640 Turns und damit 66,7 % der Fälle ein Negationswort im Kontext des Kommunikationsverbs.

Die Analyse von Adjazenzmustern und damit einhergehend die Analyse der Gebrauchskontexte von Kommunikationsverben und n-Grammen hat also die erste Hypothese bestätigt, nach der die kommunikative Gattung der Regierungspressekonferenz eine epistemische Gattung ist, eine Gattung also, in der Wissen verhandelt und die Möglichkeit der Erkenntnis selbst zum Thema gemacht wird. Dabei hat sich gezeigt, dass sprachliche Praktiken der Nicht-Vergabe von Information einen erheblichen Stellenwert haben.

6.3 Selektives Kodieren: Modellierung

Ziel des selektiven Kodierens ist es nun, die als salient identifizierte epistemische Dimension der kommunikativen Gattung der Regierungspressekonferenz präziser zu fassen, um ihrer soziokulturellen Bedeutsamkeit auf die Spur zu kommen. Hierfür sollen Sprachhandlungen, mit denen Informationsaustausch realisiert wird, gemessen, die Muster ihrer Abfolge klassifiziert und zu einem Modell zusammengefasst werden, das eine Aussage über die Funktion von Regierungspressekonferenzen ermöglicht: ein stochastisches Modell der Verteilung sprachlicher Muster in einer zweizügigen Frage-Antwort-Sequenz.

Für die Messung sprachlicher Muster wurde auf eine Taxonomie von Sprachhandlungen zurückgegriffen, die für handlungstheoretisch fundierte Ansätze im Bereich Deutsch als Fremdsprache entwickelt wurde (Glaboniat et al. 2005). Diese Taxonomie ordnet den einzelnen Sprechakttypen auch sprachliche Muster zu. Diese sprachlichen Muster wurden für die vorliegende Analyse in ein maschinell verarbeitbares Format gebracht und für die vorliegenden Daten behutsam ergänzt (vgl. Scharloth 2016). Von dieser Taxonomie wurden die Muster der Verständigungssicherung, des Informationsaustauschs, der

Bitte um Stellungnahme, der Rechtfertigung, der Meinungsäußerung, des Ausdrucks von Konsens und Dissens, der Realisierbarkeit sowie der Beurteilung von Zuständen, Ereignissen und Handlungen in die Analyse einbezogen.

Da im Rahmen dieses Beitrags lediglich der Forschungsprozess illustriert werden kann, wäre es übertrieben, von der Entwicklung einer Theorie zu sprechen. Stattdessen sollen im Folgenden jene Aspekte, die auf den Ebenen offenen und axialen Kodierens als relevante Merkmale von Regierungspressekonferenzen identifiziert wurden, modelliert werden. Der Prozess der Modellierung (Stachowiak 1973) besteht aus den Operationen der Abgrenzung, also der Nichtberücksichtigung irrelevanter Original-Objekte (hier: Selektion durch Korpusbildung), der Reduktion als dem Weglassen von Objektdetails, die dem pragmatischen Zweck nicht dienlich sind (hier: Analysefokus auf Muster des Informationsaustauschs), der Dekomposition als eine Zerlegung in einzelne Segmente, der Aggregation, verstanden als eine Vereinigung von Segmenten zu einem Ganzen, und der Abstraktion, also der Begriffs- und Klassenbildung als einem Beitrag zur Theoretisierung des Forschungsgegenstands (vgl. Scharloth 2016).

Die Dekomposition als Operation der Zerlegung in einzelne Segmente erfolgte mittels eines Mapping der Sprachhandlungstypen auf die jeweiligen Turns mit Hilfe der sprachlichen Muster aus Glaboniat et al. (2005). Die für den Modellierungsprozess zentrale Operation der Aggregation wurde in mehrere Schritte aufgeteilt. Zunächst wurde anhand der Distribution der einzelnen Sprechakttypen ermittelt, welche Sprachhandlungen für die jeweiligen Turns typisch sind. Im Anschluss wurde die typische Abfolge von Sprechakten innerhalb eines Turns mittels der durchschnittlichen Position ihres Auftretens berechnet, um schließlich typische turnübergreifende Adjazenzbeziehungen zwischen Mustern, die für die Praktiken des Informationsaustauschs konstitutiv sind, zu ermitteln. Das Ergebnis ist in Abbildung 3.2 visualisiert.

Im Hinblick auf die Musterabfolgen innerhalb der Turns muss zunächst festgestellt werden, dass die relativen Positionen im Turn der Regierungsvertreter/innen, die allesamt um 0.5 liegen, ein Indiz dafür sind, dass es für das Gesamtmuster keine feste, sondern vielmehr sehr variable Abfolge der Sprachhandlungstypen gibt. Eine Clusteranalyse könnte hier womöglich differenziertere Ergebnisse zutage fördern. Im Turn der Journalist/innen hingegen wird eine Ordnungstendenz sichtbar: Ausgehend von der Darstellung von Fakten (als gegeben, wahr darstellen / als nicht gegeben, nicht wahr darstellen / identifizieren, benennen) oder dem Zweifel an ihnen (Zweifel ausdrücken) werden Meinungen formuliert und spekuliert (Glauben ausdrücken, als möglich darstellen) und im Anschluss nach Informationen, Meinungen und Wissen gefragt, ehe schließlich auf die Zuständigkeit und Kompetenz Bezug genommen und ggf. nach Rechtfertigung verlangt wird.

Journalist_in
Informationen erfragen (2331, 0.68)
von Eventuellem sprechen (1934, 0.72)
nach Zuständigkeit fragen (1765, 0.8)
nach Fähigkeit fragen (1732, 0.8)
als gegeben, wahr darstellen (Affirmation) (832, 0.37)
als selbstverständlich darstellen (824, 0.64)
sich vergewissern (750, 0.88)
identifizieren, benennen (663, 0.56)
Glauben ausdrücken (587, 0.63)
Rechtfertigung verlangen (550, 0.82)
als nicht gegeben, nicht wahr darstellen (Negation) (544, 0.47)
zustimmen, beipflichten (486, 0.74)
auf etwas aufmerksam machen (441, 0.78)
Meinungen erfragen (387, 0.77)
als möglich darstellen (351, 0.74)
versichern (348, 0.61)
Wissen ausdrücken (316, 0.57)
Äußerungen wiedergeben (298, 0.7)
Vermutungen ausdrücken (209, 0.59)
Zweifel ausdrücken (194, 0.41)
nach Wissen fragen (187, 0.78)

Regierung
von Eventuellem sprechen (3221, 0.63)
Nicht-Zuständigkeit ausdrücken (3211, 0.64)
Glauben ausdrücken (2947, 0.59)
Zuständigkeit, Kompetenz ausdrücken (2840, 0.57)
Fähigkeit ausdrücken (2827, 0.57)
als gegeben, wahr darstellen (Affirmation) (2775, 0.57)
identifizieren, benennen (2567, 0.61)
als nicht gegeben, nicht wahr darstellen (Negation) (1745, 0.6)
versichern (1602, 0.6)
Vermutungen ausdrücken (1383, 0.56)
als selbstverständlich darstellen (1354, 0.56)
korrigieren (1344, 0.61)
auf etwas aufmerksam machen (1217, 0.52)
an etwas erinnern (1204, 0.51)
Wissen ausdrücken (1063, 0.58)
zustimmen, beipflichten (894, 0.56)
verneinen (884, 0.5)
verallgemeinern, generalisieren (831, 0.65)
ankündigen (804, 0.69)
Nichtwissen ausdrücken (771, 0.6)
widersprechen (703, 0.51)
Informationen erfragen (673, 0.59)
zugeben, eingestehen (672, 0.57)
als möglich darstellen (644, 0.58)
Zweifel ausdrücken (624, 0.57)
Überzeugung ausdrücken (605, 0.61)
Äußerungen wiedergeben (442, 0.63)

Abb. 3.2: Distribution von Sprachhandlungen in Turns von Journalisten und darauf folgenden Turns von Regierungsvertretern. In Klammern die absolute Frequenz und die 0/1-normalisierte relative Position im jeweiligen Turn.

Betrachtet man die Frequenz von Sprachhandlungstypen je Turn, so treten in den Journalist/innen-Turns das Erfragen von Informationen und das Spekulieren mit Konditionalen (von Eventuellem sprechen) am häufigsten auf, dicht gefolgt von Fragen nach Kompetenz und Autorisierung. In den Turns der Regierungsvertreter/innen ist die Rede von Eventuellem der häufigste Sprachhandlungstyp, gefolgt vom Ausdruck mangelnder Zuständigkeit bzw. fehlender Autorisierung, der häufiger vorkommt als Selbstzuschreibung der Zuständigkeit (Zuständigkeit, Kompetenz ausdrücken). Unsicherheit und probabilistische Kalküle (Glauben ausdrücken) werden auch sehr häufig verwendet. Auch hier zeigt sich also das leichte Übergewicht der expliziten Nichtvergabe von Information aus Gründen der Nichtzuständigkeit.

7 Fazit

Es ist diese spezifische Verschränkung von Autorität und Wissen, die in der Kombination sprachlicher Muster sichtbar wird und die soziokulturelle Bedeutung der kommunikativen Gattung ausmacht: Dass nämlich die Regierungspressekonferenz nicht allein in der Funktion der Informationsvergabe aufgeht, sondern dass sie neben der epistemischen auch eine performative Funktion hat. Indem die Regierungsvertreterinnen und -vertreter immer wieder auf die Möglichkeit und Autorisiertheit zur Vergabe von Informationen bzw. auf die fehlende Autorisierung zur Preisgabe von Informationen verweisen, erzeugen sie Verhülltes und Geheimes und rufen mit ihm auch jene Rollen und Institutionen ins Gedächtnis, die darüber entscheiden können, was geheim bleibt und was nicht. Die Regierungspressekonferenz ist damit auch als Ritual lesbar, das die *arcana imperii* mitkonstituiert und mit ihnen Macht nicht nur darstellt und in Erinnerung bringt, sondern auch erzeugt.

Diese Deutung der kommunikativen Gattung wurde mittels einer an das Vorgehen der Grounded Theory angelehnten korpuslinguistischen Methodologie erarbeitet, die versuchte, die sprachlichen Daten durch ihren wechselseitigen Bezug aufeinander als immer schon interpretierte Daten sichtbar zu machen. Um diese Bezüge herzustellen, kamen Analyseverfahren zum Einsatz, die mit korpuslinguistischen Standard-Tools nicht durchführbar gewesen wären. Sich von Daten im Forschungsprozess anleiten zu lassen, bedeutet daher auch, dass sich die Methoden an die Daten anschmiegen müssen. Ein an die Prinzipien der Grounded Theory angelehnter Forschungsprozess verlangt damit auch von Korpuslinguistinnen und Korpuslinguisten ein breites Kompetenzprofil, das sowohl kultur- und sozialwissenschaftliches Wissen, als auch fundierte Programmierkenntnisse umfasst.

Literatur

Anderson, Chris (2008): The end of theory: The data deluge makes the scientific method obsolete. *WIRED MAGAZINE* 16.07 (06/23/08). http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/ (letzter Zugriff: 15. 6. 2017).

Breuer, Franz (2010): *Reflexive Grounded Theory. Eine Einführung für die Forschungspraxis.* Unter Mitarbeit von Barbara Dieris und Antje Lettau. 2. Aufl. Wiesbaden: Springer VS.

Bundesregierung (2016): *Bundespresseamt/Geschichte und Information.* https://www.bundesregierung.de/Content/DE/StatischeSeiten/Breg/Bundespresseamt/bundespresseamt-das-amt-im-ueberblick.html (letzter Zugriff: 15. 6. 2017).

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse* (Sprache und Wissen 4). Berlin, New York: de Gruyter.

Bubenhofer, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Sven Roth, Martin Wengeler & Alexander Ziem (Hrsg.), *Handbuch Sprache in Politik und Gesellschaft* (Handbücher Sprachwissen 19), 69–93. Berlin, Boston: de Gruyter.

Bubenhofer, Noah & Joachim Scharloth (2016): Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik. In Ludwig Jäger, Werner Holly, Peter Krapp, Samuel Weber & Simone Heekeren (Hrsg.), *Sprachwissenschaft als Kulturwissenschaft: Sprache – Kultur – Kommunikation/Language – Culture – Communication* (Handbücher zur Sozial- und Kommunikationswissenschaft 43), 924–933. Berlin, Boston: de Gruyter.

Burger, Harald (1998): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer & Ralf Klabunde (Hrsg.) (2010): *Computerlinguistik und Sprachtechnologie*. 3. Aufl. Heidelberg, Berlin: Springer Spektrum.

Evert, Stefan (2009): Corpora and collocations. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics. An international handbook*, Vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2), 1212–1248. Berlin, New York: de Gruyter.

Feilke, Helmuth (2000): Die pragmatische Wende in der Textlinguistik. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (Hrsg.), *Text- und Gesprächslinguistik/Linguistics of text and conversation*, 1. Halbband (Handbücher zur Sprach- und Kommunikationswissenschaft 16.1), 64–82. Berlin, New York: de Gruyter.

Feilke, Helmuth (2003): Textroutine, Textsemantik und sprachliches Wissen. In Angelika Linke, Hanspeter Ortner & Paul R. Portmann-Tselikas (Hrsg.), *Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis* (Reihe Germanistische Linguistik 245), 209–230. Tübingen: Niemeyer.

Firth, John Rupert (1957): Modes of meaning. In *Papers in linguistics 1934–1951*, 190–215. London: Oxford University Press.

Geertz, Clifford (1973): *The interpretation of cultures: Selected essays.* New York: Basic Books.

Geertz, Clifford (1983): *Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme*. Frankfurt am Main: Suhrkamp.

Glaboniat, Manuela, Martin Müller, Paul Rusch, Helen Schmitz & Lukas Wertenschlag (2005): *Profile deutsch. Gemeinsamer europäischer Referenzrahmen. Lernzielbestimmungen, Kann-Beschreibungen, Kommunikative Mittel, Niveau A1–A2, B1–B2, C1–C2*. Berlin u. a.: Langenscheidt.

Glaser, Barney & Anselm Strauss (1967): *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine Publishing Company.

Harras, Gisela, Edeltraud Winkler, Sabine Erb & Kristel Proost (Hrsg.) (2004): *Handbuch deutscher Kommunikationsverben*. Teil 1: *Wörterbuch* (Schriften des Instituts für deutsche Sprache 10.1). Berlin, New York: de Gruyter.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Henning Bergenholtz & Joachim Mugdan (Hrsg.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984* (Lexicographica Series Maior 3), 118–129. Tübingen: Niemeyer.

Klein, Josef (2001): Gespräche in politischen Institutionen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (Hrsg.), *Text- und Gesprächslinguistik/Linguistics of text and conversation.* 2. Halbband (Handbücher zur Sprach- und Kommunikationswissenschaft 16.2), 1589–1606. Berlin, New York: de Gruyter.

Köhler, Miriam M. & Christian H. Schuster (2006): Regierungs-PR im Feld der politischen Kommunikation / Funktion und Bedeutung von regierungsamtlicher Presse- und Öffentlichkeitsarbeit. In Miriam M. Köhler (Hrsg.): *Handbuch Regierungs-PR: Öffentlichkeitsarbeit von Bundesregierungen und deren Beratern*, 13–32. Wiesbaden: Springer VS.

Manning, Christopher D. & Hinrich Schütze (2002): *Foundations of statistical natural language processing.* 5. Aufl. Cambridge, MA: The MIT Press.

Nerlich, Brigitte (1995): The 1930s – At the birth of a pragmatic conception of language. *Historiographica Linguistica* XXII (3), 311–334.

Perkuhn, Rainer & Cyril Belica (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. *Sprachreport* 22 (1), 2–8.

Popper, Karl Raimund (1966): *Logik der Forschung.* 2., erw. Aufl. Tübingen: Mohr (Siebeck).

Scharloth, Joachim (2016): Praktiken modellieren: Dialogmodellierung als Methode der Interaktionalen Linguistik. In Arnulf Deppermann, Helmuth Feilke & Angelika Linke (Hrsg.), *Sprachliche und kommunikative Praktiken* (Jahrbuch des Instituts für Deutsche Sprache 2015), 311–336. Berlin, Boston: de Gruyter.

Scharloth, Joachim, David Eugster & Noah Bubenhofer (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse & Wolfgang Teubert (Hrsg.), *Linguistische Diskursanalyse. Neue Perspektiven*, 345–380. Wiesbaden: Springer VS.

Schiller, Anne, Simone Teufel & Christine Thielen (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Arbeitspapier. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, Universität Tübingen, Seminar für Sprachwissenschaft.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. Arbeitspapier. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Stachowiak, Herbert (1973): *Allgemeine Modelltheorie.* Wien, New York: Springer.

Strauss, Anselm & Juliet Corbin (1996): *Grounded Theory: Grundlagen Qualitativer Sozialforschung.* Weinheim: Beltz/Psychologie Verlags Union.

Strübing, Jörg (2008): *Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung des Verfahrens der empirisch begründeten Theoriebildung*. 2., überarb. und erw. Aufl.. Wiesbaden: Springer VS.

Teubert, Wolfgang (2005): My version of corpus linguistics. *International Journal of Corpus Linguistics* 10 (1), 1–13.

Stefan Th. Gries

4 Korpuslinguistik und ihr Potenzial für die (amerikanische) Rechtsprechung

Abstract: In diesem Artikel diskutiere ich kurz, welche Anwendungsmöglichkeiten korpuslinguistische Methoden in der amerikanischen Rechtsprechung haben. Anhand von zwei weit bekannten, linguistisch aber zweifelhaften Entscheidungen des Obersten Gerichtshofs in den Vereinigten Staaten von Amerika zeige ich, wie die Analyse von Konkordanzen und Kollokaten es ermöglicht, den im amerikanischen Rechtssystem weit verbreiteten, aber schlecht definierten Begriff des *ordinary meaning* besser zu fassen und Urteilsbegründungen auf linguistisch sichereren Sachverstand zu basieren.

Keywords: Häufigkeit, Korpuslinguistik, Prototypen, (amerikanische) Rechtsprechung, Semantik

1 Einleitung

In der amerikanischen Rechtsprechung kann sich die Höhe einer Strafe aus einem normalen Strafmaß und strafverschärfenden Maßnahmen (sog. *sentence enhancements*) zusammensetzen.[1] Derartige strafverschärfende Tat-/Sachbestände liegen zum Beispiel vor,

- wenn körperliche Gewalt nicht nur zu Sachschäden, sondern auch zu Körperverletzungen führt;
- wenn ein Verbrechen begangen wird, während der Täter auf Bewährung ist;
- wenn der Täter ein Wiederholungstäter ist;
- wenn das Verbrechen Hasskriminalität darstellt (also zum Beispiel sexistisch, rassistisch, antisemitisch etc. motiviert ist);
- wenn das Verbrechen „in Verbindung mit einer Schusswaffe" verübt wurde.

1 Dieser Artikel basiert zu einem großen Teil auf einer Zusammenarbeit mit Brian G. Slocum (präsentiert als Gries & Slocum 2017, Vorarbeit für einen Artikel für den *Brigham Young University Law Review*).

Stefan Th. Gries, Department of Linguistics, University of California Santa Barbara, Santa Barbara, CA 93106-3100; U.S.A., E-Mail: stgries@linguistics.ucsb.edu

https://doi.org/10.1515/9783110538649-005

Besonders im letzten Fall kann die Strafverschärfung teilweise erheblich ausfallen; selbst Verbrechen, die sonst nur mit wenigen Jahren Haft bestraft werden würden, können in diesem Fall unverzüglich mit zehn Jahren Haft oder sogar deutlich mehr bestraft werden. Für die genauen Folgen ist die exakte Formulierung des Gesetzes relevant, die hier daher zitiert werden soll (18 U.S.C. 924(c)(1)(A), meine Hervorhebung):

> (A) Except to the extent that a greater minimum sentence is otherwise provided by this subsection or by any other provision of law, **any person who, during and in relation to any crime of violence or drug trafficking crime** (including a crime of violence or drug trafficking crime that provides for an enhanced punishment if committed by the use of a deadly or dangerous weapon or device) for which the person may be prosecuted in a court of the United States, **uses or carries a firearm**, or who, in furtherance of any such crime, possesses a firearm, shall, in addition to the punishment provided for such crime of violence or drug trafficking crime –
> [jede Person, die während und in Verbindung mit einer Gewalttat oder mit illegalem Drogenhandel … eine Schusswaffe verwendet oder mitführt …]
>
> (i) be sentenced to a term of imprisonment of not less than 5 years;
> (ii) if the firearm is brandished, be sentenced to a term of imprisonment of not less than 7 years; and
> (iii) if the firearm is discharged, be sentenced to a term of imprisonment of not less than 10 years.
>
> (B) If the firearm possessed by a person convicted of a violation of this subsection –
> (i) is a short-barreled rifle, short-barreled shotgun, or semiautomatic assault weapon, the person shall be sentenced to a term of imprisonment of not less than 10 years; or
> (ii) is a machinegun or a destructive device, or is equipped with a firearm silencer or firearm muffler, the person shall be sentenced to a term of imprisonment of not less than 30 years. (https://www.law.cornell.edu/uscode/text/18/924 [letzter Zugriff: 18. 4. 2017])

Dieser Gesetzestext wirft einige Fragen auf, die in Verfahren in amerikanischen Gerichten – von Bezirksgerichten über Appellationsgerichtshöfe bis zum Obersten Gerichtshof der Vereinigten Staaten – kontrovers diskutiert wurden; interessanterweise sind es Fragen, die im Prinzip nichts anderes als lexikalische Semantik sind:

- was bedeutet *carry a firearm* [eine Schusswaffe tragen]?
- was bedeutet *use a firearm* [eine Schusswaffe verwenden]?

In den folgenden Abschnitten werden diese Fragen diskutiert. Im nächsten Abschnitt werde ich zunächst einen kurzen Überblick darüber geben, welche Grundsätze das amerikanische Rechtssystem beinhaltet, wenn es um die Interpretation von Wörtern geht, bevor ich dann kurz zwei Fallstudien bespreche,

die sich auf den obigen Gesetzestext und die Interpretationen von *carry a firearm* und *use a firearm* beziehen. In jeder dieser kurzen Fallstudien werde ich zuerst die Sachlage schildern, dann erläutere ich die entsprechende Entscheidung des Obersten Gerichtshofs und wie sie entstanden ist beziehungsweise begründet wurde, bevor ich schließlich eine korpuslinguistische Sicht auf den jeweiligen Fall diskutiere.

2 Bedeutung und „normale Bedeutung"

Ein wichtiger Begriff in der amerikanischen Rechtsprechung ist *ordinary meaning*, im Folgenden als „normale Bedeutung" übersetzt. Der Oberste Gerichtshof in den Vereinigten Staaten hat argumentiert, dass Wörter, die nicht in einem Gesetz definiert werden, mit ihren normalen oder üblichen Bedeutungen verstanden werden (Smith vs. United States, No. 91-8674, 1993, https://supreme.justia.com/cases/federal/us/508/223/case.html [letzter Zugriff: 18. 4. 2017]). Dies ist u. a. dann der Fall, wenn es keine andere Quelle gibt, aus der vernünftig erschlossen werden könnte, was der Gesetzgeber im Sinn hatte, als er den Gesetzestext formulierte, oder was ein sorgfältiger Leser des Gesetzes verstehen würde (übersetzt von Watson vs. United States, No. 06-571, 2007, https://supreme.justia.com/cases/federal/us/552/74/opinion.html [letzter Zugriff: 18. 4. 2017]). Bedauerlicherweise scheint es aber allen Beteiligten und insbesondere dem Obersten Gerichtshof klar zu sein, dass selbst die gründlichsten Leser wahrscheinlich an der Mehrheit der Gesetzestexte scheitern würden. Richter Alito beurteilte die Komplexität der relevanten Gesetzestexte im Zivilfall *Perry vs. Merit Systems Protection Board* wie folgt:

> The one thing about this case that seems perfectly clear to me is that nobody who is not a lawyer, and no ordinary lawyer, could read these statutes and figure out what they are supposed to do.
>
> [Das einzige, was mir an diesem Fall ganz klar ist, ist dass niemand, der kein Rechtsanwalt ist, und auch kein/nicht einmal ein normaler Rechtsanwalt, diese Gesetzestexte lesen könnte und verstehen würde, was er zu tun hat.] (Barnes 2017, o. S.)

Aus einer etwas akademischeren Sicht zeigen Lee & Mouritsen (im Erscheinen), dass aber auch die Umsetzung der Richtlinie, dass Wörter in ihren „normalen" oder „üblichen" Bedeutungen verstanden werden sollen, durchaus problematisch ist. Wenn Richter über „normale Bedeutung" sprechen, dann definieren sie als „normal" manchmal verschiedene Punkte auf dem folgenden Kontinuum:

possible → common → most frequent (→ prototypical) → exclusive
möglich → verbreitet → am häufigsten (→ prototypisch) → exklusiv

Mit anderen Worten, in einigen Fällen haben Richter eine Bedeutung eines Wortes identifiziert, die unter bestimmten Bedingungen möglich ist, und haben diese dann als die „normale Bedeutung" deklariert. Wie Lee & Mouritsen (im Erscheinen) am berühmten Beispiel des hypothetischen „No vehicles in the park"-Warnschilds zeigen, ist es möglich, dieses Schild mit Referenzen auf verschiedene Punkte dieses Kontinuums zu interpretieren. *Vehicle* kann ein Tier bezeichnen, das ein „Träger" einer Infektion ist (eine mögliche Bedeutung), aber *vehicle* kann auch ein Fahrrad bezeichnen (eine verbreitete Bedeutung), aber die häufigste und prototypische Bedeutung ist sicherlich ein Auto mit vier Reifen und einem Verbrennungsmotor.

3 Zwei kurze Fallstudien

3.1 Muscarello vs. United States (1998)

3.1.1 Die Analyse des Obersten Gerichtshofs

In diesem Fall war die für das zu verhängende Strafmaß entscheidende Frage, was *carry a firearm* bedeutet. Frank Muscarello hatte eine Schusswaffe im verschlossenen Handschuhfach seines Fahrzeugs, während er Marihuana verkaufte und bis er verhaftet wurde. Ist das ein Sachverhalt, der durch die hervorgehobene Formulierung im oben zitierten Gesetzestext 18 U.S.C. 934(c)(1)(A) abgedeckt ist, hier leicht gekürzt „any person who during and in relation to a crime of drug trafficking carries a firearm"? Der Oberste Gerichtshof bejahte diese Frage, was zur Folge hatte, dass Frank Muscarellos Strafmaß von normalerweise ca. 10–16 Monaten auf ca. 5,5 Jahre angehoben wurde und dass diese Entscheidung Präzedenzcharakter für viele ähnlich geartete Situationen haben könnte.

Wie kam es zu dieser Entscheidung? Die Mehrheit des Obersten Gerichtshofes (fünf der neun Richter), deren Urteil von Richter Stephen Breyer geschrieben wurde, näherte sich dieser Fragestellung auf der Basis einer beachtlichen und diversen Menge an Quellen und Belegen, unter anderem die King-James-Bibel, literarische Werke wie *Robinson Crusoe* und *Moby Dick*, aber auch Zeitschriftenkorpora und Wörterbücher. (Die Minderheit des Obersten Gerichtshofes argumentierte auf der Basis eines juristischen Wörterbuchs, anderen Bibelübersetzungen, Gedichten, Zitaten aus dem Film *The magnificent seven*

und aus den Fernsehserien *MASH* und der *Sesamstraße* …) Auf dieser Basis folgert Breyer für die Mehrheit, dass die erste oder Hauptbedeutung („first, or primary meaning") von *carry* „Beförderung" („carrying as conveyance") sei, wohingegen eine andere und besondere („different, rather special") Bedeutung „(am Körper) tragen" sei („the notion of carrying upon one's person"). Die Mehrheit folgert explizit, dass „the relevant linguistic facts are that the word carry in its ordinary sense includes carrying in a car" (Mouritsen 2010: 1926).

Von besonderem Interesse für unsere Diskussion hier ist die Verwendung von Wörterbüchern und die fehlerhaften, wenn auch weit verbreitete Annahmen, die in die Verwendung von Wörterbüchern einflossen. Eine dieser Annahmen ist die, dass die Reihenfolge von Bedeutungen, die in einem Wörterbuch angegeben werden, notwendigerweise etwas über die normale/grundlegende Bedeutung eines Wortes aussagt. Breyer argumentiert:

> Consider first the word's primary meaning. The Oxford English Dictionary gives as its **first** definition „convey, originally by cart or wagon, hence in any vehicle, by ship, on horseback, etc." 2 OXFORD ENGLISH DICTIONARY 919 (2d ed. 1989); see also WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY 343 (1986) (**first** definition: „move while supporting (as in a vehicle or in one's hands or arms)"; RANDOM HOUSE DICTIONARY OF THE ENGLISH LANGUAGE UNABRIDGED 319 (2d ed. 1987) (**first** definition: „to take or support from one place to another; convey; transport").
>
> [Betrachten wir zunächst die primäre Bedeutung des Wortes. Das *Oxford English Dictionary* gibt als erste Definition „transportieren, ursprünglich mit einem Wagen, daher in jeder Art Fahrzeug, per Schiff, auf einem Pferd etc." […] vgl. auch *Webster's Third International Dictionary* (erste Definition: „bewegen, während man es gegen die Schwerkraft unterstützt (wie in einem Fahrzeug oder mit Händen oder Armen)"); *Random House Dictionary of the English Language Unabridged* […]: (erste Definition „von einem Platz zu einem anderen bewegen, befördern, transportieren")]
>
> (https://supreme.justia.com/cases/federal/us/524/125/case.html [letzter Zugriff: 18. 4. 2017])

Breyer begeht hier den Irrtum, dass er der Reihenfolge der Bedeutungen von *carry* eine Relevanz zuschreibt, die die Reihenfolge nicht hat, wie Mouritsen (2010) ausführlich diskutiert. Erstens, die Herausgeber des *Webster's Third New International Dictionary* warnen sehr deutlich:

> The system of separating by numbers and letters reflects something of the semantic relationship between various senses of a word. It is only a lexical convenience. It does not evaluate senses or establish an enduring hierarchy of importance among them. The best sense is the one that most aptly fits the context of an actual genuine utterance.
>
> [In diesem System werden durch Zahlen und Buchstaben Aspekte der semantischen Relationen zwischen verschiedenen Bedeutungen eines Wortes angezeigt. Dieses System ist nur der Bequemlichkeit geschuldet. Weder bewertet es Bedeutungen noch klassifiziert es

> sie nach Wichtigkeit. Die beste Bedeutung ist die, die am besten in den entsprechenden Kontext einer tatsächlichen Verwendung/Äußerung passt.] (zit. nach Mouritsen 2010: 1930)

Außerdem erklären sie, dass die Reihenfolge der Bedeutungen historisch motiviert sei: die erste Bedeutung sei die, von der man wisse, dass sie zuerst im Englischen verwendet wurde und dass manchmal eine willkürliche Bedeutungsreihenfolge gewählt wurde – sie erklären allerdings nicht, wie eine historische von einer willkürlichen Reihenfolge unterschieden werden kann.

Zweitens, die Herausgeber des *Oxford English Dictionary* (zit. nach Mouritsen 2010: 1933) geben ebenfalls an, dass ihre Bedeutungsreihenfolge historisch motiviert sei – wichtiger sind dennoch die Tatsachen, dass (i) sie im Prinzip nur zwei Bedeutungen unterscheiden, genau die beiden, die Breyer diskutiert, aber (ii) darüber hinaus suggerieren, dass die erstere immer weniger verwendet wird, was im Umkehrschluss darauf hinaus läuft, dass die zweite Bedeutung tatsächlich die hauptsächliche ist.

Drittens, das *Random House Dictionary* (zit. nach Mouritsen 2010: 1935) listet seine Bedeutungen nicht auf historischer Grundlage auf, sondern typischerweise („generally") auf der Grundlage der Häufigkeiten, mit denen Bedeutungen verwendet wenden. Wie Mouritsen (2010) zeigt, sind die Reihenfolgen jedoch nicht notwendigerweise akkurat, da sie von Heuristiken und Verzerrungen in der Wahrnehmung von Bedeutungen beeinträchtigt sein können; außerdem basieren die Häufigkeiten, die die Herausgeber typischerweise heranziehen, nicht auf Korpusdaten, sondern auf den gesammelten Belegen des Herausgeberteams.

Eine weitere problematische Annahme ist die, dass die „zentrale Bedeutung" auf der Basis von Etymologien – ebenfalls aus Wörterbüchern – erschlossen werden kann. Diese Vorgehensweise ist offensichtlich problematisch, da es in keinster Weise offensichtlich ist, dass die Etymologie eines Wortes auch nur irgendetwas mit einer zeitgemäß zentralen Bedeutung zu tun hat: Dezember ist eben nicht (mehr) der zehnte Monat. (Mouritsen diskutiert noch einige weitere Probleme, die die Verlässlichkeit der Wörterbuchanalysen für den Obersten Gerichtshof beeinträchtigen, die ich hier allerdings nicht betrachten werde, weil sie weniger Relevanz für die spätere korpuslinguistische Diskussion haben.)

3.1.2 Eine korpuslinguistische Perspektive auf Muscarello vs. United States

In Anbetracht der Tatsache, dass der Oberste Gerichtshof die Wichtigkeit der „zentralen Bedeutung" erkannt hat, stellt sich die Frage, wie man sich diesem

Konzept korpuslinguistisch nähern kann, und die einfachste Operationalisierung ist natürlich die über Korpushäufigkeit. Ein Ansatz der bessere Resultate erzielen sollte, sofern das verwendete Korpus repräsentativer ist als die Beispielsammlung des *Random House Dictionarys*, wovon meines Erachtens ausgegangen werden kann. Betrachten wir kurz zwei Korpusanalysen, eine von Mouritsen (2010) und eine von Lee & Mouritsen (im Erscheinen):

Mouritsen (2010) extrahiert eine Zufallsstichprobe von *carry* als Verb aus dem Corpus of Contemporary American English (COCA, Davies 2008–). Die Mehrheit der Beispiele in seiner Stichprobe sind Bedeutungen von *carry*, die nicht tatsächlich wörtlich „transportieren" oder „tragen" bedeuten, aber Mouritsen findet auch eine größere Anzahl an Verwendungen von *carry* innerhalb phrasaler Verben oder Partikelverben. Insgesamt findet Mouritsen, dass 5 % der Verwendungen von *carry* die Beförderungsbedeutung haben, während 29 % die Bedeutung „am Körper tragen" haben; vergleicht man nur diese beiden Bedeutungen, so ist das Verhältnis der beiden Bedeutungen 15 % zu 85 %.

Mouritsen führte auch noch eine weitere Konkordanzanalyse durch, dieses Mal eine, in der *carry* in Verbindung mit einem aus einer kleinen Menge an Waffenbegriffen (*firearm* [Schusswaffe], *gun* [Schusswaffe], *handgun* [Handfeuerwaffe], *rifle* [Gewehr], *pistol* [Pistole]) stammenden Wort verwendet wird. Für die klar identifizierbaren Verwendungen ist das Resultat sogar noch eindeutiger als das oben genannte kollokationsunspezifische Ergebnis: 1 % der Bedeutungen beziehen sich auf „Beförderung", wohingegen sich 64 % auf „am Körper tragen" beziehen.

Die zweite Fallstudie, Lee & Mouritsen (im Erscheinen), ist eine Analyse von *carry* im News on the Web Corpus (NOW, Davies 2013), einem dynamischen Korpus mit zurzeit ca. 4,3 Milliarden Wörtern. 271 Verwendungen von *carry* mit den gleichen Kollokaten wie oben wurden analysiert. Von diesen 271 Verwendungen beinhalten 5 die Bedeutung „Beförderung" und 104 die Bedeutung „am Körper tragen".

Fassen wir zusammen. Ich denke, jeder Korpuslinguist und Semantiker würde uneingeschränkt zugeben, dass die Prototypikalität/Zentralität einer Bedeutung B nicht zwingenderweise bedeutet, dass B auch die häufigste Bedeutung sein muss; auch der Umkehrschluss, dass die häufigste Bedeutung *zwingenderweise* die zentrale Bedeutung ist, gilt nicht. Außerdem ist es offensichtlich, dass die höhere Häufigkeit einer Bedeutung nicht beweist, dass es diese Bedeutung ist, die der Gesetzgeber (hier, der Kongress der Vereinigten Staaten) im Sinne hatte, als das Gesetz formuliert wurde. Nichtsdestotrotz sollte klar sein, dass in Abwesenheit expliziterer Definitionen/Formulierungen in Gesetzestexten oder anderer empirischer Belege eine derartig häufigere Bedeutung eher die „zentrale Bedeutung" ist, die der Oberste Gerichtshof ja als so zentral für die Rechtsprechung anführte, als eine derartig seltenere.

3.2 Smith vs. United States (1993)

3.2.1 Die Analyse des Obersten Gerichtshofs

Der zweite Fall ist wahrscheinlich noch erstaunlicher, was die Interpretation des relevanten Wortes angeht. Es geht wieder um den o. g. Paragraphen zu strafverschärfenden Bedingungen. John A. Smith wurde verhaftet, nachdem er einem Zivilbeamten anbot, eine halbautomatische Maschinenpistole vom Typ MAC-10 gegen Kokain einzutauschen. Der Fall ging über mehrere Instanzen, bis er vor dem Obersten Gerichtshof landete: Smith und seine Anwälte gestanden, dass die Schusswaffe während („during") und in Beziehung zu („in relation to") dem Drogenvergehen auftauchte – die relevante Frage war allerdings, ob das Eintauschen einer Schusswaffe gegen Drogen ein Fall von „*use a firearm* during and in relation to a drug trafficking crime" sei, so dass strafverschärfende Tatbestände vorlagen. Sowohl das Bezirksgericht in Südflorida als auch der 11. Appellationsgerichtshof in Atlanta, Georgia, waren der Ansicht, dass der oben zitierte Gesetzestext das Eintauschen einer Schusswaffe abdecke – wenn der Gesetzgeber gemeint hätte, dass die Verwendung einer Schusswaffe nur die Verwendung als Waffe, nicht aber als Tauschware/Währung, bedeuten sollte, dann hätte der Gesetzgeber dies entsprechend formulieren müssen; da er das nicht getan hatte, würde die generelle Bedeutung von *use a firearm* gelten. Da jedoch der 9. Appellationsgerichtshof in San Franciso, California, in einem ähnlichen Fall anderer Ansicht war, landete der Fall vor dem Obersten Gerichtshof.

Der Oberste Gerichtshof schloss sich dem Bezirksgericht und dem 11. Appellationsgerichtshof an. Zur Begründung wurden wie in *Muscarello vs. United States* u. a. Wörterbücher herangezogen. Zum einen wurde argumentiert, dass Smith und seine Anwälte mit ihrem Einspruch/ihrer Petition nur recht bekommen könnten, wenn der fragliche Gesetzestext nicht „use a firearm" sondern „use a firearm as a weapon" lauten würde, was er nicht tut. Zum anderen und da Rechtsprechung im Zweifelsfall auf der „normalen Bedeutung" der im Gesetzestext verwendeten Ausdrücke basieren sollte, zitierte die Mehrheit des Obersten Gerichtshofes zwei Wörterbücher: *Webster's Dictionary* definiert *to use* als „to convert to one's service" („zu seinem Dienste verwenden") oder „to employ" („verwenden"/„gebrauchen"), *Black's Law Dictionary* offeriert eine ähnliche Definition, und auf der Grundlage dieser Definitionen argumentierte der Oberste Gerichtshof,

> Smith's handling of the MAC-10 in this case falls squarely within those definitions. By attempting to trade his MAC-10 for the drugs, he "used" or "employed" it as an item of barter to obtain cocaine; he "derived service" from it because it was going to bring him the very drugs he sought.

[Die Verwendung der MAC-10 Maschinenpistole durch Smith [den Antragsteller] hier fällt unter genau diese Definitionen. Mit dem Versuch, seine MAC-10 gegen Drogen einzutauschen, „verwendete“ oder „gebrauchte“ er sie als einen Tauschgegenstand, um Kokain zu erhalten, er „profitierte“ davon, weil sie ihm genau die Drogen beschafft hätte, die er zu erhalten versuchte.]
(https://supreme.justia.com/cases/federal/us/508/223/case.html [letzter Zugriff: 18. 4. 2017])

Die Minderheit des Obersten Gerichtshofs, insbesondere Richter Scalia, kritisierte diese Entscheidung scharf, da der Kontext des Wortes *use* und seine „normale Bedeutung“ nicht angemessen berücksichtigt wurde. Scalia argumentierte, „to use an instrumentality ordinarily means to use it for its intended purpose“ [ein Instrument zu verwenden heißt normalerweise, es für seinen intendierten Zweck zu benutzen] und „to speak of using a firearm is to speak of using it for its distinctive purpose, i. e. as a weapon“ [zu sagen, dass man eine Schusswaffe benutzt, bedeutet, dass man davon spricht, sie für ihren entsprechenden Zweck zu verwenden, das heißt als eine (Schuss)Waffe]. Er fügte hinzu,

[w]hen someone asks, “Do you use a cane?”, he is not inquiring whether you have your grand-father’s silver-handled walking stick on display in the hall; he wants to know whether you walk with a cane.

[Wenn jemand fragt „Gebrauchst Du einen Spazierstock?”, dann fragt er nicht, ob man den Spazierstock seines Großvaters mit einem silbernen Griff im Flur ausstellt; er fragt, ob man mit einem Stock läuft.]
(https://supreme.justia.com/cases/federal/us/508/223/case.html [letzter Zugriff: 18. 4. 2017])

Seiner Ansicht nach waren die Worte *as a weapon* damit implizit im Gesetzestext enthalten. Die Mehrheit lehnte diese Argumentation u. a. mit einer Begründung ab, die zitiert werden muss, um wirklich zu verstehen wie – aus meiner Sicht – problematisch und widersprüchlich sie ist (meine Hervorhebung):

It is one thing to say that the ordinary meaning of “uses a firearm” includes using a firearm as a weapon, since that is the intended purpose of a firearm and the example of “use” *that most immediately comes to mind*. But it is quite another to conclude that, as a result, the phrase also excludes any other use.

[Es ist eine Sache zu sagen, dass die „normale Bedeutung“ von „uses a firearm“ [„verwendet eine Schusswaffe“] die Bedeutung beinhaltet „verwendet eine Schusswaffe als eine Waffe“, da das der normalerweise beabsichtigte Verwendungszweck einer Schusswaffe ist und dies die Bedeutung ist, die einem als erstes einfällt. Aber es ist etwas ganz anderes, daraus zu schlussfolgern, dass die Phrase „use a firearm“ andere Bedeutungen ausschließt.]
(https://supreme.justia.com/cases/federal/us/508/223/case.html [letzter Zugriff: 18. 4. 2017])

Das Interessante an dieser Begründung ist ihr Verhältnis zur Diskussion in Abschnitt 2. Auf der einen Seite hat der Oberste Gerichtshof wie viele andere Instanzen auch die „Richtlinie" der „normalen Bedeutung" angenommen; diese Richtlinie oder dieses Prinzip wurde nicht umsonst von Slocum (2015) als das fundamentalste aller Prinzipien für juristische Interpretation bezeichnet. Auf der anderen Seite wird die Meinung der Minderheit – insbesondere von Richter Scalia – nicht einmal direkt adressiert: Scalia hatte nicht argumentiert, dass die „normale Bedeutung" von *use a firearm* „eine Schusswaffe als (Schuss)Waffe" *beinhaltet* („includes"), sondern dass das die „normale Bedeutung" *ist*. Noch wichtiger ist jedoch, dass die Mehrheit sogar zugesteht, dass dies die Bedeutung ist, die einem als erstes / sofort einfallen würde (!) – man kommt nicht umhin sich zu fragen, welche Definition von „normaler Bedeutung", die nicht mit dem Kriterium „that most immediately comes to mind" kompatibel ist oder sogar übereinstimmt, die Mehrheit denn annimmt.

3.2.2 Eine korpuslinguistische Perspektive auf Smith vs. United States

Um zu testen, was Korpusdaten zu der Interpretation von *use a firearm* beitragen können, haben Gries & Slocum (2017) die Resultate einer Korpusanalyse vorgestellt, deren Design der von Mouritsen (2011) ähnelt. Wir verwendeten ein Skript in der Programmiersprache R (R Core Team 2016), um aus dem 2012–2015 Update für COCA (Davies 2008–) alle Beispiele zu extrahieren, die dem folgenden Suchmuster entsprachen:

- das Lemma *use*, wenn es als Verb getaggt ist („^v");
- gefolgt von einem Determiner oder einem Possessivpronomen („^(dd[12]|at1?|appge)$");
- möglicherweise gefolgt von einem Adjektiv („jj[rt]?");
- gefolgt von einem Substantiv („nn[12]").

Wir erhielten 21,2T Treffer, von denen 145 Beispiele waren, deren folgender Kontext eines von mehreren Waffen-Substantiven enthielt (*gun* und seine Derivate *rifle*, *firearm*, *pistol* und *weapon*). Dann prüften wir, wie viele dieser 145 Beispiele von *use [some weapon]* die Bedeutung hatten „als Tauschware verwenden" – das Resultat war null, nicht ein einziges Beispiel hatte die von der Mehrheit des Obersten Gerichtshofs angenommene „normale Bedeutung".

Natürlich könnte ein Skeptiker nun argumentieren, dass selbst wenn *use [some weapon]* nicht „normalerweise" mit der Tauschbedeutung verwendet wird – um es milde auszudrücken – die Tauschbedeutung doch zumindest ein (integraler?) Bestandteil der *use* + DO Mini-Konstruktion ist. Daher prüften wir

noch 159 zusätzliche zufällig ausgewählte Beispiele von *use* mit einem beliebigen konkreten direkten Objekt und zählten, wie viele dieser Beispiele von *use + [some concrete object]* die Bedeutung hatten „als Tauschware verwenden"; das Ergebnis war das Gleiche: kein einziges Beispiel hatte diese Bedeutung.

Unsere Daten bestreiten offensichtlich nicht die Tatsache, dass *use + [some concrete object]* (ob nun eine Waffe oder etwas anders) die Bedeutung „als Tauschgegenstand verwenden" *haben kann* – sie zeigen jedoch, dass diese Phrase diese Bedeutung *normalerweise nicht hat*; „als Tauschgegenstand verwenden" ist eine mögliche Bedeutung, aber offensichtlich weder eine verbreitete noch die häufigste noch die prototypische und schon gar nicht die exklusive, was mit der Minderheitenargumentation insbesondere von Richter Scalia übereinstimmt.

Interessanterweise lässt sich als eine Art Epilog festhalten, dass in einem zumindest ähnlichen späteren Fall – *Watson vs. United States (2003)*, nach einer Berufung gegen ein Urteil des 5. Appellationsgerichtshofs in New Orleans, Louisiana – der Oberste Gerichtshof anders entschied, wenn auch nicht unbedingt aus Gründen, die mit der obigen korpuslinguistischen Analyse einhergehen. In *Watson vs. United States* etablierte der Oberste Gerichtshof u. a. eine Unterscheidung in Bezug darauf, wer die Waffe verwendet. In *Smith vs. United States* wurde argumentiert, dass Smith, der eine Schusswaffe weggab und dafür Drogen erhalten wollte, eine Waffe im Sinne von 18 U.S.C. 924(c)(1)(A) verwendet – in *Watson vs. United States* wurde argumentiert, dass Watson, der Drogen weggab und dafür eine Schusswaffe erhalten wollte, nicht eine Waffe im Sinne von 18 U.S.C. 924(c)(1)(A) verwendet. Während die Begründung für diese Entscheidung nichts mit der Definition des Wortes *use* in 18 U.S.C. 924(c)(1)(A) zu tun hatte (sondern sich eher auf die Bedeutung von *carry* bezog und die Zeitspanne während derer Watson im Besitz der Schusswaffe war), ist Richter Ginsburgs zusätzliche Stellungnahme (*concurrence*) aus linguistischer Sicht interessanter und es wert, hier in Gänze und als diesen Abschnitt abschließende Bemerkung zitiert zu werden:

> It is better to receive than to give, the Court holds today, at least when the subject is guns. Distinguishing, as the Court does, between trading a gun for drugs and trading drugs for a gun, for purposes of the 18 U.S.C. § 924(c)(1) enhancement, makes scant sense to me. I join the Court's judgment, however, because I am persuaded that the Court took a wrong turn in Smith v. United States, 508 U. S. 223 (1993), when it held that trading a gun for drugs fits within § 924(c)(1)'s compass as "us[e]" of a firearm "during and in relation to any … drug trafficking crime." For reasons well stated by Justice Scalia in his dissenting opinion in Smith, 508 U. S., at 241, I would read the word "use" in § 924(c)(1) to mean use as a weapon, not use in a bartering transaction. Accordingly, I would overrule Smith, and thereby render our precedent both coherent and consistent with normal usage. Cf. Henslee v. Union Planters Nat. Bank & Trust Co., 335 U. S. 595, 600 (1949) (Frankfurter, J., dissenting) ("Wisdom too often never comes, and so one ought not to reject it merely because it comes late.").

> [Der Gerichtshof argumentiert heute „es ist besser zu erhalten als zu geben", zumindest wenn es um Schusswaffen geht, und unterscheidet zwischen eine Schusswaffe gegen Drogen und Drogen gegen eine Schusswaffe einzutauschen, eine Unterscheidung, die mir wenig sinnvoll erscheint. Ich schließe mich jedoch trotzdem der Mehrheit an, weil ich überzeugt bin, dass das Gericht in *Smith vs. United States* eine Fehlentscheidung getroffen hat, als es argumentiert hat, dass eine Schusswaffe gegen Drogen einzutauschen unter die Definition von „Verwendung einer Schusswaffe während und in Verbindung mit einem Drogenhandelsverbrechen" in § 924(c)(1) fällt. Aus den von Richter Scalia in seiner Minderheitsmeinung wohl formulierten Gründen würde ich die Bedeutung von „verwenden" in § 924(c)(1) als „als eine (Schuss)Waffe verwenden" ansehen und nicht als „als Tauschgegenstand verwenden". Dementsprechend würde ich Smith überstimmen/annullieren und damit unseren Präzedenzfall kompatibel mit normalem Sprachgebrauch machen. Vgl. *Henslee v. Union Planters Nat. Bank & Trust Co.*, 335 U. S. 595, 600 (1949), Richter Frankfurters Minderheitenmeinung („Allzu oft kommt Weisheit gar nicht und daher sollte man sie nicht ablehnen, nur weil sie spät kommt.")]
>
> (https://supreme.justia.com/cases/federal/us/552/74/concurrence.html [letzter Zugriff: 18. 4. 2017])

4 Abschließende Bemerkungen

Dass die Sprachwissenschaft der Jurisprudenz wichtige Dienste leisten kann, steht wohl außer Frage, wie an Gebieten wie der forensischen Linguistik schon lange deutlich ist. Jedoch geht der potenzielle Nutzen der Linguistik über Autoren- oder Dialektidentifikation und vergleichbare Anwendungen hinaus. Wie ich hoffentlich zeigen konnte, ist selbst die linguistische Semantik und die korpuslinguistische Methode zu ihrer Erforschung potenziell hoch relevant. Auch wenn die oben diskutierten Fallstudien aus korpuslinguistischer Sicht alle methodologisch höchst einfach waren und über eine extrem einfache Annotation von Konkordanzzeilen nicht hinausgingen, so sind weitere und komplexere, herausforderndere Anwendungen denkbar, wie z. B. die Fragen,

- ob und/oder wie sich die Bedeutung eines Wortes in einem Gesetzestext seit der Verabschiedung eines Gesetzes geändert hat (was z. B. durch technologische Entwicklungen in der Kommunikation hoch relevant geworden ist, die zu neuen Definitionen oder Interpretationen von *Postgeheimnis* und *Privatsphäre* führen mussten; ein konkretes Beispiel in den U.S.A. ist, ob die Bedeutung des Verbs *harbor* heutzutage die gleiche ist wie 1952, als der Gesetzgeber dieses Verb einer Formulierung im *Immigration and Nationality Act* hinzufügte, eine zentrale Frage in *United States vs. Costello [2012]*);
- ob bestimme Denotate von einem linguistischen Ausdruck abgedeckt werden (vgl. das obige Beispiel von „no vehicles in the park");
- ob zwei Ausdrücke synonym sind oder nicht und, falls nicht, wie sie sich funktional (semantisch, pragmatisch, konnotativ, etc.) unterscheiden; etc.

Außerdem ist selbst die oben diskutierte Art von Frage – was ist die „normale Bedeutung“ eines linguistischen Ausdrucks? – nicht immer oder ausschließlich durch die Operationalisierung „die häufigste Verwendung“ zu beantworten, denn prototypische Bedeutungen können, müssen aber natürlich nicht, die häufigsten sein, und die Korpuslinguistik stellt andere Forschungsmethoden zur Verfügung (Typenfrequenz und Informationsgehalt von Kollokaten, Dispersion u. a.), die in solchen Fällen nützlich sein können. (Natürlich gibt es aber auch Fragestellungen, die mit korpuslinguistischen Methoden schwer(er) zu untersuchen sind und daher ggf. experimentelle oder andere Verfahren erfordern. Beispielsweise wurde ich einmal gebeten, linguistisch zu belegen, dass der Ausdruck *six or fewer* null beinhaltet (und nicht nur sechs bis eins). Es ist nicht offensichtlich, dass man selbst in größeren Korpora genug hinreichend eindeutige Beispiele finden könnte, die einen Richter oder eine Jury überzeugen würden, so dass ich für diese Frage keine Korpusdaten verwendet habe.)

Bedauerlicherweise ist es allerdings immer noch der Fall, dass (in den U.S.A. zumindest) Richter oft davon ausgehen, dass sie qua ihres Berufes, indem sie zweifelsohne viel mit Sprache umgehen, gut genug als linguistische Experten fungieren können und unwillig sind, linguistische Expertisen von Gutachtern zuzulassen. Gerade aus einer gebrauchsbasierten Perspektive ist diese Sicht überaus problematisch: Gerade, *weil* Richter viele Gesetzestexte, Präzedenzfälle und andere Gutachten lesen, die ja allesamt wenig dem „normalen Sprachgebrauch’ entsprechen, ist ihr Sprachwissen eben nicht repräsentativ für den „normalen Sprachgebrauch“! Genauso wenig wie Anwälte und Richter nicht mit der Wimper zucken würden, wenn es darum geht, Expertengutachen in vielen Wissenschaftsfeldern zuzulassen, so sollten sie hier vielleicht auch ihres eigenen blinden Flecks gewahr werden und erkennen, dass der multidimensionale und probabilistische mentale Sprachraum auch etwas ist, was Expertenwissen verlangt – die Konsequenzen für Angeklagte sind zu schwerwiegend, als dass man aus vielleicht falscher Eitelkeit die Art von Expertise, die Linguisten im allgemeinen und Korpuslinguisten im speziellen anbieten können, einfach verleugnen sollte.

Literatur

Barnes, Robert (2017): Hearing first arguments as member of the Supreme Court, Gorsuch jumps right in. The Washington Post, Online-Ausgabe vom 17. 4. 2017. https://www.washingtonpost.com/politics/courts_law/hearing-first-arguments-as-member-of-the-supreme-court-gorsuch-jumps-right-in/2017/04/17/96856404-2392-11e7-a1b3-faff0034e2de_story.html?utm_term=.bffe5711be5e (letzter Zugriff: 30. 1. 2017).

Davies, Mark (2016): Corpus of Contemporary American English. URL: http://corpus.byu.edu/coca/ (letzter Zugriff: 30. 1. 2017).

Gries, Stefan Th. & Brian G. Slocum (2017): Ordinary meaning and corpus linguistics. (Vorträge gehalten auf dem Law and Corpus Linguistics workshop 2017, Brigham Young University, Provo, UT)

Lee, Thomas R. & Stephen C. Mouritsen (Im Erscheinen): Judging ordinary meaning. *Yale Law Review*. Available at SSRN: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937468> (letzter Zugriff: 30. 1. 2017).

Mouritsen, Stephen C. (2011): The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review* 1915–2010. Available at SSRN: <https://ssrn.com/abstract=1753333> (letzter Zugriff: 30. 1. 2017).

R Core Team (2016): R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. URL: <https://www.R-project.org>

Slocum, Brian G. (2015): *Ordinary meaning: A theory of the most fundamental principle of legal Interpretation*. Chicago, IL: The University of Chicago Press.

Stefanie Dipper und Sarah Kwekkeboom

5 Historische Linguistik 2.0

Aufbau und Nutzungsmöglichkeiten der historischen Referenzkorpora des Deutschen

Abstract: In den letzten Jahren entstanden eine Reihe von Korpora zu historischen Sprachstufen, darunter auch die Korpora, die zum Verbund der historischen Referenzkorpora des Deutschen zählen: ReA (Referenzkorpus Altdeutsch), ReM (Referenzkorpus Mittelhochdeutsch), ReF (Referenzkorpus Frühneuhochdeutsch), ReN (Referenzkorpus Mittelniederdeutsch/Niederrheinisch) und ReDI (Referenzkorpus Deutsche Inschriften). Im folgenden Beitrag werden die Referenzkorpora vorgestellt und die Herausforderungen beschrieben, die sich bei der Transkription und Annotation der Texte ergaben. Im zweiten Teil werden Nutzungsmöglichkeiten der Referenzkorpora für die historische Linguistik beschrieben. Der Schwerpunkt der Darstellung liegt auf ReM, das online* zugänglich ist und für das eine detaillierte Dokumentation vorliegt.

Keywords: Referenzkorpus, Altdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Mittelniederdeutsch, Inschriften, Transkription, Annotation, Korpussuche, ANNIS

1 Design der Korpora: Textauswahl

Die Referenzkorpora zu historischen Sprachstufen des Deutschen bilden die Grundlage für ein sprachstufenübergreifendes Textkorpus, das sowohl histo-

Anmerkung: Wir danken den Mitarbeiterinnen und Mitarbeitern in den Einzelprojekten der Referenzkorpora, die an der Entstehung der Korpora beteiligt waren, über die hier berichtet wird. Wir danken insbesondere Birgit Herbers und Klaus-Peter Wegera für wertvolle Hinweise und Korrekturen. Die Arbeit an diesem Beitrag wurde unterstützt durch die DFG (Geschäftszeichen DI-1558/1 und DI-1558/5).

* Alle in diesem Beitrag genannten URLs wurden am 29. 7. 2017 abgefragt.

Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, E-Mail: dipper@linguistics.rub.de
Sarah Kwekkeboom, Germanistisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, E-Mail: sarah.kwekkeboom@rub.de

https://doi.org/10.1515/9783110538649-006

rischsynchrone als auch diachrone Recherchemöglichkeiten bietet. Die Sprachstufen sind zu großen Teilen nach den gleichen Maßstäben diplomatisch transkribiert und digital erfasst und nach einheitlichen Kriterien mit Wortarten und flexionsmorphologischen Informationen annotiert. Zudem sind die Texte mit weitestgehend gleichen Metainformationen in Form von Headerdaten ausgestattet. Die erarbeiteten Daten werden über das browserbasierte Korpustool ANNIS (Krause & Zeldes 2016) für Recherchen für die Forschergemeinde frei zugänglich gemacht.

Im Jahr 2009 starteten zeitlich versetzt die Referenzkorpusprojekte „Altdeutsch, von den Anfängen bis 1050 (ReA)“ sowie „Mittelhochdeutsch, 1050–1350 (ReM)“. Das „Referenzkorpus Frühneuhochdeutsch, 1350–1650 (ReF)“ folgte 2012, „Mittelniederdeutsch/Niederrheinisch, 1200–1650 (ReN)“ 2013. Mit etwas anderen Parametern und einer anderen Ausgangslage die Textgrundlage betreffend, startete 2014 das Referenzkorpus „Deutsche Inschriften (ReDI)“.[1]

Grundsätzlich richten sich alle Korpusprojekte, bis auf ReDI, bei der Textauswahl nach einer dreiteiligen Gliederung: Zeitraum, Sprachraum und Textart. Alle drei Bereiche sollten so abgedeckt werden, dass eine repräsentative Darstellung der Sprachstände der jeweiligen Sprachstufe gewährleistet ist. Im folgenden Abschnitt sollen Textauswahl und Korpusdesign der einzelnen Teilkorpora kurz vorgestellt werden.

1.1 Referenzkorpus Altdeutsch, von den Anfängen bis 1050 (ReA)

Das Referenzkorpus Altdeutsch[2] erfasst sämtliche althochdeutschen und altniederdeutschen Texte und deckt damit den Zeitraum von ca. 750–1050 ab. Insgesamt enthält ReA die fünf größeren Texte althochdeutscher und altsäch-

---

1 An den einzelnen Projekten waren bzw. sind beteiligt:

ReA: Karin Donhauser (Humboldt-Universität Berlin), Rosemarie Lühr (Friedrich-Schiller-Universität Jena), Jost Gippert (Goethe-Universität Frankfurt)

ReM: Klaus-Peter Wegera, Stefanie Dipper (Ruhr-Universität Bochum), Thomas Klein, ab 2013 Claudia Wich-Reif (Rheinische Friedrich-Wilhelms-Universität Bonn)

ReF: Klaus-Peter Wegera, Stefanie Dipper (Ruhr-Universität Bochum), Hans-Joachim Solms (Martin-Luther-Universität Halle-Wittenberg), Ulrike Demske (Universität Potsdam)

ReN: Ingrid Schröder (Universität Hamburg), Robert Peters (Westfälische Wilhelms-Universität Münster)

ReDI: Klaus-Peter Wegera (Ruhr-Universität Bochum), Birgit Herbers (Johannes Gutenberg-Universität Mainz)

2 http://www.deutschdiachrondigital.de/

sischer Zeit (Isidor, Tatian, Otfrid, Notker und Heliand) sowie eine Vielzahl kleinerer Textdenkmäler beider Sprachstufen, die den Editionen von Elias Steinmeyer (*Die kleineren althochdeutschen Sprachdenkmäler*, Berlin 1916) und Elis Wadstein (*Kleinere altsächsische sprachdenkmäler*, Norden/Leipzig 1899) entnommen sind.

Als besondere Herausforderung und Alleinstellungsmerkmal für ReA stellte sich der enge Bezug der ältesten deutschsprachigen Literatur zu den lateinischen Vorlagen heraus; parallel überlieferte lateinische Texte und Textpassagen sowie Glossen wurden ebenfalls erfasst und vollständig annotiert.

ReA umfasst ca. 650.000 tief annotierte Wortformen und ist seit 2014 über ANNIS durchsuchbar.[3]

1.2 Referenzkorpus Mittelhochdeutsch, 1050–1350 (ReM)

Das Referenzkorpus Mittelhochdeutsch[4] enthält die hochdeutschen Sprachdenkmäler von ca. 1050 bis 1200 weitestgehend vollständig und von ca. 1200 bis 1350 in strukturierter Auswahl. Die Texte wurden im ReM (anders als im ReA) diplomatisch transkribiert, d. h. Editionen wurden für die Erfassung der Textzeugen nur unterstützend verwendet. Insgesamt bietet ReM etwa 3,1 Mio. digitalisierte Wortformen, davon 2,5 Mio. tief annotiert, die sich auf fünf Teilkorpora unterschiedlicher Ausrichtung verteilen (s. auch Klein & Dipper 2016: 2 ff.):

(i) Das „Frühmittelhochdeutsch-Korpus“ (1050–1200) enthält alle überlieferten Textzeugen der älteren mittelhochdeutschen Zeit bis zum Ende des 12. Jahrhunderts, die bis dahin für sprachhistorische Untersuchungen nur unzureichend erschlossen waren. Es weist ca. 210 mittelhochdeutsche Quellen mit annähernd 1 Mio. digitalisierter und ca. 815.000 annotierter Wortformen auf.

(ii) Das „Korpus zur Mittelhochdeutschen Grammatik“ (MiGraKo) umfasst 102 Texte bzw. Textausschnitte im Umfang von je rund 12.000 Wortformen. Diese wurden (im Rahmen des Projekts „Mittelhochdeutsche Grammatik“) nach korpuslinguistischen Kriterien ausgewählt (Näheres zur Korpusauswahl s. Wegera 2000), nach der Handschrift diplomatisch transkribiert, präeditiert und interpungiert, komplett lemmatisiert und hinsichtlich der flexionsmorphologischen Kategorien annotiert. Die ersten beiden Zeiträume des MiGraKo (1070–1150 und 1150–1200) sind Teil des

3 https://korpling.german.hu-berlin.de/annis3/ddd/
4 https://www.linguistics.ruhr-uni-bochum.de/rem/

Frühmittelhochdeutschkorpus. Im ReM wurden diese Texte, wo möglich, auf 20.000 Wortformen erweitert.

(iii) Bei den Prosagroßtexten handelt es sich um zwölf Langtexte, die im MiGraKo bereits als annotierte Teiltexte mit einem Umfang von je ca. 12.000 Wortformen vorlagen. Diese wurden komplett digitalisiert und bis zur Grenze von 24.000 Wortformen tief annotiert.

(iv) Das strukturierte „Erweiterungskorpus A“ zur Mittelhochdeutschen Grammatik umfasst weitere 41 Handschriften des 13./14. Jahrhunderts, die im Prinzip nach den gleichen korpuslinguistischen Kriterien wie die Texte des MiGraKo ausgewählt wurden. Durch diese Verbreiterung der Quellenbasis spiegelt das mittelhochdeutsche Referenzkorpus die Diversität der mittelhochdeutschen Überlieferung und der regionalen Schreibsprachen noch angemessener wider. Alle Texte des „Erweiterungskorpus A“ wurden komplett oder in Ausschnitten von ca. 12.000 Wortformen erfasst, jeweils 6.000 Wortformen wurden annotiert. Insgesamt wurden hier rund 430.000 Wortformen erfasst, von denen rund 220.000 Wortformen tief annotiert sind.

(v) Das „Erweiterungskorpus B“ spiegelt ausschließlich den mitteldeutschen-Sprachraum wider, dieser ist in der mittelhochdeutschen Gesamtüberlieferung vor dem 14. Jahrhundert erheblich schlechter vertreten als der oberdeutsche. Das „Erweiterungskorpus B“ umfasst insgesamt annotierte 171.000 Wortformen.

Die Auswahl eines Textes erfolgte nach den bereits erwähnten Parametern Zeit, Raum und Überlieferungsform, wobei hier lediglich Prosa, Vers oder Urkunde unterschieden werden.

ReM ist seit 2016 über ANNIS verfügbar.[5]

1.3 Referenzkorpus Frühneuhochdeutsch, 1350–1650 (ReF)

Während ReM nur in Teilen ein balanciertes Korpus ist (nämlich ab 1200, davor wurden alle Textzeugen aufgenommen), wurde die Textauswahl für ReF einer weitaus strengeren und strukturierteren Ordnung unterzogen, d. h. nur ein Auszug der überlieferten und greifbaren Textzeugen wurde hier aufgenommen. Die Textauswahl erfolgte dabei weiterhin nach den Parametern Zeit, Raum und Überlieferungsform, hinzu kam aber, dass neben Handschriften ab 1450 auch gedruckte Texte vorhanden sind, die gleichwertig zu den Handschriften eine

5 Zugang über https://www.linguistics.ruhr-uni-bochum.de/rem/

Daseinsberechtigung im Korpus haben. Während die ersten Zeiträume von ReF (1350–1450) folglich nur Handschriften aufweisen, wurden für den Zeitraum 1450–1550 gleichermaßen Handschriften und Drucke bearbeitet. Die Textmenge verdoppelt sich daher in diesen Abschnitten. Ab 1550 wurden dann ausschließlich Drucke bearbeitet.

Tab. 5.1: Das Feld IV-Cc mit vier Textfeldern.

		Cc Hessisch	
1500–1550	Handschrift	IV-Cc-T1 Rinck: Widerlegung (kirchlich-theologisch)	IV-Cc-T2 Gerstenberg Landeschronik (chronikalische u. Berichtstexte)
	Druck	IV-Cc-T1 Dryander Arzneibuch (Realientext)	IV-Cc- T2 Halsgerichtsordnung Karl V. (Rechts- und Geschäftstext)

Innerhalb des Korpus wurden die verschiedenen und für das Frühneuhochdeutsche charakteristischen Textsorten angemessen berücksichtigt. Die Texte wurden in einem induktiv-pragmatischen Verfahren sechs Textbereichen zugewiesen (Rechts- und Geschäftstexte, chronikalische und Berichtstexte, Realientexte, unterhaltende Texte, kirchlich-theologische Texte/Bibeln, erbauliche Texte). Nach Möglichkeit sollte jede Kombination („Feld") von Zeit- und Sprachraum neben Vers- und Prosatexten eine Auswahl dieser Textsorten bieten. Auch wenn dies aus verschiedenen Gründen nicht immer möglich war, so sollte ein solches Feld im Idealfall etwa einen Inhalt wie in Tabelle 5.1 aufweisen.

Ähnlich wie in ReM wurde auch in ReF ein bereits existierendes Korpus integriert. Das „Bonner Frühneuhochdeutsch-Korpus" (Solms & Wegera 1998) entstand ab 1972 in der Forschungsstelle „Frühneuhochdeutsch" der Universität Bonn. Auf Grundlage dieses Korpus wurden verschiedene Bände der Grammatik des Frühneuhochdeutschen erarbeitet. Alle 40 Quellen werden jetzt nach den Standards der Referenzkorpusprojekte neu aufbereitet und nachannotiert.

Alle Texte in ReF werden vollständig oder mit bis zu 20.000 Wortformen diplomatisch erfasst und mit 12.000 Wortformen tief annotiert. Anders als in den anderen Referenzkorpusprojekten werden für ReF neben der morphosyntaktischen und flexionsmorphologischen Annotation Teile des Korpus auch syntaktisch analysiert. Mit dem Tool @nnotate (Brants & Skut 1998) wird in der Arbeitsstelle Potsdam eine strukturierte Auswahl von 21 Texten zusätzlich auch syntaktisch annotiert. Auch diese Annotationsebene wird über ANNIS abrufbar und recherchierbar sein.

Letztlich soll das Korpus insgesamt 202 Texte mit 4 Mio. digitalisierten Wortformen enthalten, davon ca. 2,4 Mio. morphologisch und 430.000 syntaktisch annotiert.

1.4 Referenzkorpus Mittelniederdeutsch/Niederrheinisch, 1200–1650 (ReN)

Genau wie ReF wurde auch das Referenzkorpus Mittelniederdeutsch/Niederrheinisch nach Feldern aufgebaut, die die Parameter Zeit, Raum und Textart enthalten. Auch das ReN-Projekt strukturiert den dritten Parameter Textart nach eigenen Maßstäben und benennt hier sieben „Felder der Schriftlichkeit“ (Verwaltung, Recht, Wissensvermittlung, Religion (geistliche Literatur), Weltliche Literatur, private Schriftlichkeit, Inschriften). Jedem der sieben Felder werden relativ feststehend verschiedene Textsortengruppen und Quellengattungen zugewiesen (Peters & Nagel 2014: 168 f.).

Ähnlich wie auch in ReM finden in ReN verschiedene bereits bestehende, digitale Korpora Platz im Referenzkorpusprojekt. Die Texte werden auch hier den bestehenden Standards angepasst und neu annotiert. Es handelt sich dabei um drei in Münster bestehende digitale Korpora (Peters & Nagel 2014: 170 ff.):

(i) Das Projekt „Atlas spätmittelalterlicher Schreibsprachen des niederdeutschen Altlandes und angrenzender Gebiete (ASnA)“ erfasste insgesamt 5.547 Texte (zumeist Urkunden). Die 44 Ortspunkte des Projekts decken dabei mit etwa 20 Texten pro Ortspunkt aus dem 13. bis 15. Jahrhunderts einen großen Teil des niederdeutschen Sprachgebiets ab.

(ii) Aus den rund 1.000 Texten des Projekts „Mittelniederdeutsch in Lübeck (Historisches Digitales Textarchiv) (MiL)“ finden sich 50 Lübecker Urkunden, eine Stadtrechts-Handschrift, das Schiffsrecht, eine Chronik, eine Historien-Bibel sowie 12 Drucke im Referenzkorpus wieder.

(iii) Aus einem Korpus-Projekt zur Dokumentation niederdeutscher Sprachzeugnisse aus Westfalen mit dem Titel „Niederdeutsch in Westfalen (Historisches Digitales Textarchiv) (NiW)“ wurden für ReN ebenfalls Texte herangezogen. Es handelt sich hier um niederdeutsche (westfälische) Sprachdenkmäler vom 9. bis zum 19. Jahrhundert. In ReN sind daraus ausschließlich Texte enthalten, die auf Grundlage der Originale digitalisiert wurden und deren Umfang etwa 20.000 Wortformen beträgt.

Insgesamt soll das Referenzkorpus Mittelniederdeutsch/Niederrheinisch letztlich 180 Texte umfassen (bei den Urkunden werden jeweils mehrere zusammengefasst und als ein Text betrachtet) und insgesamt ca. 3,6 Mio. Wortformen aufweisen (Schröder 2014: 154).

1.5 Referenzkorpus Deutsche Inschriften (ReDI)

Das Referenzkorpus Deutsche Inschriften ist das kleinste der aktuell fünf Referenzkorpusprojekte, die Textgrundlage ist hier eine besondere. ReDI erfasst alle deutschen Inschriften bis 1650 aus der Editionsreihe „Deutsche Inschriften". Zu Beginn der Förderungszeit waren 95 Bände (91 Druckausgaben, 4 digitale Ausgaben) erschienen, mittlerweile liegen 100 Bände (94 Druckausgaben, 6 digitale Ausgaben) vor. Bei den „Deutschen Inschriften" handelt es sich um ein interakademisches epigraphisches Publikationsunternehmen, Ziel ist die Dokumentation und Aufbereitung von Inschriften des Frühmittelalters bis zur Frühen Neuzeit. Zu Projektbeginn befanden sich von den 95 Bänden bereits 33 Bände in DIO (Deutsche Inschriften Online: Digitalisierung und Online-Bereitstellung der Inschriftenbände, Digitale Akademie Mainz[6]), aktuell sind es 50 Bände. ReDI beschäftigt sich dabei mit allen deutschsprachigen und nicht-kopialen Inschriften bis 1650. Auch wenn viele Textbände bereits in digitaler Form zur Verfügung gestellt werden konnten, mussten sie noch an die Standards der Referenzkorpusprojekte angepasst werden. So wurden beispielsweise die Inschriftentexte der Editionen mit den vorhandenen Abbildungen abgeglichen, im besten Fall waren die Abbildungen auch über DIO einsehbar und vergrößerbar oder in anderer Weise digital vorhanden, die Abbildungen in den gedruckten Editionen waren häufig zu klein, um alle Buchstaben mit letzter Sicherheit erkennen zu können. Abbildungen waren zudem in den frühen Bänden der Reihe nur in kleinerer Auswahl vorhanden.

ReDI wird etwa 300.000–500.000 annotierte Wortformen umfassen und über ANNIS verfügbar gemacht werden (Herbers 2016).

2 Transkription und Annotation

2.1 Transkription

Um historische Korpora für sprachwissenschaftliche Untersuchungen umfassend nutzen zu können und linguistische Merkmale eindeutig darzustellen, ist eine diplomatische Transkription, d. h. eine genaue Abbildung der Zeichen unerlässlich.

Für die Projekte ReM, ReF, ReN und ReDI wurden zum Abgleich in der Regel Abbildungen des zugrundeliegenden Originals verwendet. Bei verschollenen oder zerstörten Handschriften, Drucken oder Inschriften konnte allerdings nur

6 www.inschriften.net

auf mehr oder weniger handschriftengetreue Abdrucke des 19. Jahrhunderts zurückgegriffen werden. Sollte ein Transkript nicht handschriftengetreu transkribiert bzw. Editionen verwendet worden sein, so ist dies in jedem einzelnen Fall in den Metadaten des Textes (s. u. 2.3) vermerkt. Für die Erfassung der Texte in ReA wurde jeweils auf die beste verfügbare Edition einer Handschrift zurückgegriffen, die Texte dieses Projektes werden in einer „philologisch gesicherten und möglichst handschriftennahen Form bereitgestellt“ (Donhauser 2015: 37).

Eine handschriftengetreue Transkription stellt die Bearbeiter grundsätzlich vor Probleme im Umgang mit gegenwärtig nicht mehr verwendeten Schrift- und Sonderzeichen, Abkürzungen und bei der Verarbeitung von Marginalien. Die Digitalisierung der Texte sollte in einem einfachen, standardisierten Format (reines Textformat, UTF8) erfolgen, damit sie möglichst lange und unabhängig von verwendeten Fonts maschinenlesbar bleiben.

Auf Grundlage der Transkriptionen des MiGraKo wurde in den Projekten ReM, ReF und ReDI ein umfangreiches, projektinternes Handbuch erstellt, das die Wiedergabe verschiedenster Sonderzeichen und Markierungen reguliert, sowie Richtlinien für Kommentare, nach denen die Bearbeiter Unklarheiten, Auffälligkeiten o. ä. notieren können. Für die Kodierung der Sonderzeichen werden fast ausschließlich ASCII-Zeichen eingesetzt, z. B. wird für die Kodierung des langen S (*ſ*) das optisch ähnliche $-Zeichen verwendet.

Als Beispiele für den Umgang mit Sonderzeichen sind in Abbildung 5.1 einige Auszüge aus dem Handbuch angeführt. Das Transkriptionshandbuch wird mit der Veröffentlichung von ReF online einsehbar sein.

Beispiel aus Hs.	**Transkript**	**Beschreibung**
	x\y (hier v\o)	Buchstabe als Superskript. Zu beachten ist, dass nur Vokalmodifikatoren als Superskripte transkribiert werden – etwa <i\e>, <u\o>. Alle anderen superskribierten Buchstaben werden als hochgestellt aufgefasst und mit % codiert.
	,	Kürzung von <er>, <r>, <ir>, <re>, <ri>. Unabhängig von der graphischen Realisation in der Handschrift als <’>, <^{s}>, <^{z}>.
	4	Kürzung von auslautend <rum>. Bsp.: tuo4 = tuorum.
	+R …@R	Rubrizierung. Bsp.: +R De patientia @R.

Abb. 5.1: Auszug aus dem Transkriptions-Handbuch.

2.2 Annotation

Die Referenzkorpora sind auf mehreren Ebenen annotiert, u. a. mit Lemmata und grammatischer Information. Um den Annotationsprozess zu vereinfachen, nehmen die Annotationen Bezug auf Worteinheiten, die den modernen Regeln der Wortschreibung entsprechen. So können beispielsweise Wortteile miteinander verbunden werden (z. B. abgetrenntes *ge*-Präfix bei Partizipien: *gelouben* statt originalem *ge␣louben* ‚glauben') oder zusammengeschriebene Wörter getrennt (z. B. Krasis, z. B. *wilt␣u* statt originalem *wiltu* ‚willst du'). Solche Fälle werden entsprechend markiert und sind in ANNIS sichtbar und abfragbar.

Weiterhin werden Partikelverben in Fernstellung durch Markierungen „*1" und „*2" einander zugewiesen[7] (*fo bricht****2*** *nicht meer dor'Inn abe****1*** ‚so bricht nichts weiter davon ab'), oder Pronominaladverbien in ReF in Kontaktstellung zusammengeschrieben (*darvf saz d' cune degen* – statt original *dar␣vf* – ‚darauf saß der kühne Held') (s. u. 2.4).

Die grammatische Annotation erfolgte bei ReA über die Tools ELAN (Brugman & Russel 2004) und EXMARaLDA (Schmidt 2001), die Texte in ReM wurden mit Hilfe älterer Programme erstellt, die bereits zu Annotation des MiGraKo verwendet wurden (Petran et al. 2016). Seit ReF (darunter fallen also auch ReN und ReDI) wurde mit dem in Bochum entwickelten webbasierten Annotationstool CorA (Bollmann et al. 2014) gearbeitet. Als Beispiel für die Annotation in CorA wird in Abbildung 5.2 ein Ausschnitt aus ReF in der Benutzeroberfläche gezeigt.

Abb. 5.2: CorA-Benutzeroberfläche mit einem Ausschnitt aus ReF.

7 Diese Information ist in ANNIS aktuell nicht abfragbar.

Die Annotation mit Wortarten (*part of speech*, POS) erfolgt auf dem im Rahmen der Projektgruppen ReA und ReM erarbeiteten STTS-basierten Annotations-Tagset HiTS (Historisches Tagset, Dipper et al. 2013). Eine Besonderheit ist hier die doppelte Annotation, die zwischen allgemeiner und belegspezifischer Wortart unterscheidet. Das belegspezifische Tag ist in der Regel eine genauere Definition des allgemeinen Tags, z. B. kann mhd. *vil* ‚viel' ein Adjektiv (ADJ) sein, diese Information wird im allgemeinen Tag, dem „Lemma-Tag", abgelegt. Je nach Position und Verwendung kann *vil* nachgestellt (ADJN), prädikativ (ADJD) oder attributiv, vorangestellt (ADJA) sein. Diese zusätzliche Information findet sich im belegspezifischen POS-Tag (s. Abb. 5.2).

Besonders dienlich ist die Möglichkeit der doppelten POS-Annotation, wenn die Wortart im Beleg vom Lemma abweicht, z. B. *daz* ***brinnent*** *ole* ‚das brennende Öl'. Hier wird das Partizip Präsens von *brinnen* adjektivisch verwendet. Das Tagset bietet die Möglichkeit, die Wortart des zugrundeliegenden Vollverbs anzuführen (VVPS) und daneben die in diesem spezifischen Beispiel verwendete Wortart (ADJA). Entsprechend kann so auch bei der Lemmatisierung das Vollverb *brinnen* angegeben werden.

Bestimmte Flexionsmerkmale sind nur unterspezifiziert annotiert worden. So wird der Modus nur dort angegeben, wo er ausdrucksseitig an der handschriftlichen Wortform markiert ist (Klein & Dipper 2016: 19). Ähnlich ist es bei genusschwankenden Substantiven. Sofern das Genus nicht eindeutig durch ein flektiertes Adjektiv oder einen flektierten bestimmten Artikel angegeben ist, wird diese Information nicht vergeben bzw. unterspezifiziert mit „*" annotiert. Ebenso wird bei der Vergabe eines Genus bei Pluralbelegen verfahren und auf die Genusangabe verzichtet, z. B. kann ein Plural wie *menſchen* annotiert sein als „*.Nom.Pl" (s. Abb. 5.2).

Bei der Lemmatisierung wird ebenfalls unterschieden zwischen dem allgemeinen und dem belegspezifischen Lemma, meist sind diese aber identisch. Die Gestaltung der Lemmaansätze und die Bestimmung der grammatischen Kategorien richtet sich meist nach dem jeweils entsprechenden Ansatz in den verwendeten Wörterbüchern, die als Lemmaregister dienten (Klein & Dipper 2016: 12). Für ReM wurde hier das Mittelhochdeutsche Wörterbuch nach Lexer (1872–1878) verwendet, in ReF und ReDI liegt das Deutsche Wörterbuch nach Jacob und Wilhelm Grimm (1854–1971) zugrunde. Im ReN-Projekt wurde in Ermangelung einer ausreichenden Quelle ein Lemmaregister erstellt.

2.3 Headerinformationen (Metadaten)

Jeder einzelne Text in den Referenzkorpora ist mit umfangreichen Metainformationen, den sogenannten Headerinformationen, ausgestattet. Über diese

Informationen kann z. B. ein Text bzw. ein Überlieferungsträger genau identifiziert werden. Viel wichtiger ist aber die Erweiterung der Recherchemöglichkeiten für den Benutzer. Ein Teil der Metadaten ist sprachstufenübergreifend nach den gleichen Kriterien definiert, so dass sich diese Informationen später bei der Suche als Filter nutzen lassen. Beispiele für diese Informationen sind: Sprachlandschaften, Jahrhunderthälften oder Textbereiche.

Grundsätzlich lassen sich die Metadaten in allen Referenzkorpora in Gruppen zusammenfassen: Die Benutzer erhalten Informationen zum ursprünglichen Text (Lokalisierung und Datierung), zur verwendeten Handschrift (Aufbewahrungsort und Signatur, Lokalisierung, Datierung, Schreiberinformationen etc.) und es finden sich Informationen zum verwendeten Textauszug, Zustand, Vorlagen und Bearbeiter.

Über ANNIS lassen sich die textspezifischen Headerinformationen einzeln ansehen und durchsuchen. Abbildung 5.3 zeigt die Ansicht eines Beispielheaders aus ReM in ANNIS.

metadata doc M010-N1

Name	Value
abbr_ddd	Alkuin
abbr_mwb	Alkuin
annotation_by	Bochum
collation_by	Bonn, Bochum
corpus	ReM I
date	-
digitization_by	Bonn
edition	Friedrich Wilhelm (Hg.), Denkmäler deutscher Prosa des 11. und 12. Jahrhunderts, Abteilung A: Text; Abteilung B: Kommentar (Münchener Texte 8), München 1914/16 (Nachdruck in einem Band München 1960 [Germanistische Bücherei 3]), Nr. 9, A: S. 33-37, B: S. 70-79.
extent	46r,01-48r,24
extract	-
genre	P
language	mhd
language-area	alemannisch
language-region	westoberdeutsch
language-type	oberdeutsch
library	München, Staatsbibl.
library-shelfmark	Clm 7637
medium	Handschrift
notes-annotation	-

metadata doc M010-N1

Name	Value
notes-annotation	-
notes-manuscript	_
notes-transcriptic	-
online	http://www.handschriftencensus.de/15457
place	-
pre_editing_by	Loevenich (Bonn)
proofreading_by	Bochum
reference	Hs.: Blatt (r/v), Zeile
reference-second	Wilhelm 1960: Nummer, Zeile
text	Alkuins Traktat 'De virtutibus et vitiis'
text-author	-
text-language	-
text-place	-
text-source	Übersetzung
text-type	Didaxe
time	12,1
topic	Religion
annis:doc	M010-N1

Abb. 5.3: Header-Fenster in ANNIS, Text M010-N1.

2.4 Ausgewählte Probleme bei der Transkription und Annotation

Die Erarbeitung sprachstufenübergreifender Korpora stellt die Bearbeiter in vielerlei Hinsicht vor verschiedene Probleme auf allen Bearbeitungsebenen, einige dieser Probleme sollen an dieser Stelle exemplarisch vorgeführt werden.

Für die Transkription gilt beispielsweise, dass Kürzungen oftmals ambig sein können und ihre Auflösung nicht generell vorhergesagt werden kann. Kürzungen werden daher in der Transkription nicht aufgelöst, sondern möglichst handschriftennah wiedergegeben. Ein Beispiel hierfür ist die Kürzung von <er>, <r> etc., die im Transkript jeweils nur durch den Apostroph wiedergegeben werden (s. Abb. 5.1).

Ein weiteres Beispiel ist die Zusammen-/Getrenntschreibung von Pronominaladverbien. Im Neuhochdeutschen gelten diese als univerbiert, werden also in der Standardsprache generell zusammengeschrieben. In den älteren Sprachstufen findet sich aber nicht selten Distanzstellung, in der der präpositionale und der pronominale Teil durch andere Wörter getrennt auftreten. Stehen die beiden Teile in Kontaktstellung, also adjazent, findet man sowohl Getrennt- wie auch Zusammenschreibung. Um eine einheitliche Analyse dieser Fälle zu ermöglichen, wurde im MiGraKo und ReM entschieden, generell den pronominalen vom präpositionalen Teil zu trennen. Die beiden Teile werden dann getrennt voneinander als PAVD bzw. PAVAP analysiert. Die Lemmaform zeigt hingegen die beiden Teile als zusammengehörig an, s. (1) (s. auch unten 3.1.2).

(1) ***Dar␣vmbe***. *Daz dv an dem guten menſchen ſiheſt*
‚darum, dass du an dem guten Menschen siehst‘

Transkription	POS-Tag	Morpho-Tag	Lemma	Lemma-Tag
Dar	PAVD	–	dâr/+umbe	AVD
umme	PAVAP	–	umbe/dâr+	AP

Anders wurde bei der Analyse der Pronominaladverbien in ReF verfahren. Aufgrund der Nähe zum Neuhochdeutschen geht man auch hier schon von einer Univerbierung aus. Bei Zusammenschreibung werden die beiden Teile also nicht voneinander getrennt und bei getrenntgeschriebener Kontaktstellung werden sie miteinander verbunden. In Distanzstellung werden sie durch eine Markierung im Transkript als zusammengehörig markiert, s. (2) (zur Bedeutung der geschweiften Klammern, s. u.).

(2) *Das Zimmer* ***da*1*** *er* ***inne*2*** *liegt.*
‚Das Zimmer, in dem er liegt‘

Transkription	POS-Tag	Morpho-Tag	Lemma	Lemma-Tag
da*1	PAVD	–	{dar} [GD00565]	AVD
er				
inne*2	PAVAP	–	darin [GD00616]	AP

Als letztes Beispiel soll der Umgang mit (mehrteiligen) Partikelverben in ReF angeführt werden. Bei Partikelverben mit einem abgetrennten Verbzusatz werden beide Teile zunächst als zusammengehörig markiert, dann erhält das Simplex als Lemma das volle Verbkompositum, POS- und Lemma-Tag werden normal als entsprechendes Verb markiert. Der adverbiale Teil erhält als POS-Tag PT-KVZ, als Lemma-Tag AVD, das Lemma ist entsprechend nur das Adverb/die Präposition, s. (3).[8]

(3) *sie* ***machen*2*** *den Mund* ***auf*1***

Transkription	POS-Tag	Morpho-Tag	Lemma	Lemma-Tag
machen*2	VVFIN	3.Pl.Präs.Ind	aufmachen [GA06637]	VV
den				
Mund				
auf*1	PTKVZ	–	auf [GA06002]	AVD

Bei getrennten Partikelverben mit zwei oder mehr abgetrennten Partikeln wird anders verfahren: Der Verbteil erhält nur ein kursorisches Lemma, das des Simplex. Damit es nicht fälschlicherweise bei einer möglichen Suche nach dem Simplex angezeigt wird, steht es in geschweiften Klammern. Lemma-Tag und POS-Tag erhalten die Bezeichnung der entsprechenden Verbart. Die abgetrennten Verbzusätze (PTKVZ) erhalten als Lemma das Verb/Kompositum und als Lemma-Tag AVD, s. (4)

(4) *sie* ***machen*2*** *den Mund* ***auf*1*** *und* ***zu*1***

Transkription	POS-Tag	Morpho-Tag	Lemma	Lemma-Tag
machen*2	VVFIN	3.Pl.Präs.Ind	{machen} [GM00007]	VV
den				
Mund				
auf*1	PTKVZ	–	aufmachen [GA06637]	AVD
und				
zu*1	PTKVZ	–	zumachen [GZ10140]	AVD

3 Nutzung der Korpora

In diesem Kapitel sollen verschiedene Aspekte der Nutzung von Korpora beleuchtet werden. Zunächst wird beschrieben, wie im Korpus nach bestimmten Phänomenen gesucht werden kann (3.1). Die Ergebnisse der Suche können ex-

8 Die Beispiele (3) und (4) sind konstruiert.

portiert werden, um sie gegebenenfalls mit Tools wie Excel weiter zu verarbeiten (3.2). Alternativ bietet das Korpussuchtool eine einfache Frequenz-Analyse an (3.3).

3.1 Korpusabfragen

Reich annotierte Korpora wie die Referenzkorpora bieten vielfältige Suchmöglichkeiten für die historische Linguistik. Je nachdem, ob das Korpus nur gelegentlich für einfache Suchen zu Rate gezogen oder ob es für systematische, komplexe Suchen genutzt wird, bietet ReM unterschiedliche Abfrageportale, die im Folgenden vorgestellt werden.[9]

Alle Anfragen werden letztlich vom Korpussuchtool ANNIS (Krause & Zeldes 2016) ausgewertet, das auch die Ergebnisse anzeigt. Die Abfrageportale unterscheiden sich also nur auf der Seite der Nutzereingabe.[10]

3.1.1 Für Gelegenheitsnutzer: Belegsuche

Eine typische Nutzung historischer Korpora ist die **Suche nach Belegen** für eine bestimmte Wortform in einem Zeit- und Sprachraum. Solche Anfragen beziehen sich also nur auf den Text und Metadaten und nicht auf die Annotationen. ReM bietet dafür eine maßgeschneiderte Suchmaske, die „Vereinfachte ANNIS-Suchmaske“. Abbildung 5.4 zeigt die Maske mit einer beispielhaften Suchanfrage nach dem Lemma *minne*. Außerdem ist angegeben, aus welchen Perioden („Entstehungszeit“; hier: 12. Jahrhundert) und Sprachgebieten (hier: Oberdeutsch) die Texte stammen, in denen gesucht werden soll. Das grüne Fenster mit detaillierten Angaben zu den Sprachgebieten erscheint bei Mauskontakt mit dem Fragezeichen neben dem Schlüsselwort. Zusätzlich könnten die Texte auf bestimmte Themenbereiche (wie „Alltag“, „Religion“) eingegrenzt werden.

Die Ergebnisse werden in einem neuen Fenster mit der ANNIS-Oberfläche angezeigt. Für eine detaillierte Beschreibung dieser Oberfläche bietet ANNIS ein Tutorial, das über den Reiter „Help/Examples“ zugänglich ist. Die Suchergebnisse werden im Reiter „Query Result“ angezeigt (s. Abb. 5.5).

---

9 Die Abfrageportale sind von der ReM-Website aus verlinkt: https://www.linguistics.ruhr-uni-bochum.de/rem/

10 ANNIS bietet eine weitere Abfragemöglichkeit, den „Query Builder“. Damit können Anfragen mit graphischer Unterstützung erstellt werden.

Abb. 5.4: Vereinfachte ANNIS-Suchmaske für Gelegenheitsnutzer (Screenshot der ReM-Website).

Abb. 5.5: Suchmaske und Anzeige der Treffer in ANNIS (Screenshot der ReM-Website).

Insgesamt ergab unsere Anfrage 315 Treffer in 56 verschiedenen Texten. Die Treffer sind durchnummeriert und die Sigle des Textes mit der Fundstelle wird genannt (beim ersten Treffer: „M010-N1“). Zu jedem Treffer wird ein kleiner Ausschnitt des Kontextes angezeigt, das gesuchte Wort *minne* ist dabei rot markiert. Möchte man weitere Informationen zu einem bestimmten Treffer, kann man durch einen Klick auf das „i“-Symbol oberhalb der Textstelle die oben beschriebenen Metadaten zum Text erhalten. Durch einen Klick auf das „+“Symbol unterhalb der Textstelle werden die Annotationen der Textstelle aufgeklappt, z. B. zeigt Abbildung 5.5, dass im ersten Treffer *minne* als Nomen Appelativum („NA“) annotiert ist und hier im Nominativ Singular („Nom.Sg“) steht.

Das vereinfachte Suchfenster kann auch für die **Suche nach einfachen linguistischen Phänomenen** genutzt werden, wie z. B. Lautwandel-Phänomenen wie der neuhochdeutschen Diphthongierung (*hûs* > *Haus*). Dafür müssen bei der einfachen Suchmaske die relevanten Lemmata bereits bekannt sein, man sucht also beispielsweise nach dem Lemma *hûs*. Auf dieses Beispiel gehen wir im Abschnitt 3.3 näher ein.

Außerdem erlaubt das Suchinterface, nach Wortanfängen bzw. Wortenden zu suchen, indem die entsprechenden Optionen angeklickt werden. Beispielweise kann nach Lemmata gesucht werden, die auf *-hûs* enden oder mit *hûs-* beginnen. Die häufigsten im ReM-Korpus belegten Lemmata sind in Tabelle 5.2 aufgelistet. Wie solche Frequenz-Tabellen in ANNIS erzeugt werden können, wird im Abschnitt 3.3 erläutert.

Tab. 5.2: Die häufigsten Lemmata auf *-hûs* bzw. *hûs-*.

Lemma	Frequenz	Lemma	Frequenz
hûs	1.348	*hûs*	1.348
gotes-hûs	239	*hûs-vrouwe*	121
bëte-hûs	105	*hûs-ge-nôz(e)*	55
wîg-hûs	18	*hûsen*	12
ding-hûs	13	*hûs-hêrre*	12
muos-hûs	12	*hûs-wirt*	8
olèi-hûs	12		

Bei dieser Art von Anfragen ist zu beachten, dass nicht unterschieden wird zwischen echten Affixen und „zufälligen“ Buchstabenfolgen. Beispielsweise findet die erste Suchanfrage auch das Lemma *thûs* ‚Weihrauch‘, die zweite auch das Lemma *hûse* ‚Stör‘.[11]

11 Komposita werden im Lemma durch einen Bindestrich markiert, daher kann nach solchen Bildungen gesucht werden, indem man z. B. nach *-hûs* (mit führendem Bindestrich) mit der

3.1.2 Für Experten: Suche nach komplexen Phänomenen

Das vereinfachte Suchportal erlaubt nur die Suche nach konkreten Wortformen und einer Auswahl von Metadaten. Möchte man nach weiteren Metadaten wie z. B. dem Titel eines Textes suchen[12] oder nach Annotationen wie z. B. bestimmten morphologischen Formen, so nutzt man dazu das normale ANNIS-Suchfenster, das sich im ANNIS-Portal links oben befindet (s. Abb. 5.5). In Abbildung 5.5 ist bereits eine Anfrage im Suchfenster zu sehen, die, etwas umformatiert, in (5) wiederholt wird. Diese Anfrage wurde automatisch generiert auf Basis der Anfrage, die wir in der vereinfachten Suchmaske eingegeben haben (s. Abb. 5.4).

(5)
```
lemma=/minne/ &
meta::time=/(.*12.*)/ &
meta::language-region=/(.*westober.*|.*ostober.*|.*nordober.*)/
```

Wie solche Anfragen zu formulieren sind, wird in diesem Abschnitt anhand einiger Beispiele skizziert. Für eine detaillierte Beschreibung der Anfragesprache sei auf das ANNIS-Tutorial verwiesen.

Die Anfragen werden im Suchfenster links oben eingegeben und durch einen Klick auf „Search“ gestartet. Unterhalb des Suchfensters stehen die Teilkorpora gelistet, in denen gesucht werden kann. Durch Klick bzw. Shift-Klick können ein oder mehrere Teilkorpora ausgewählt werden. Die Namen der Teilkorpora in ReM deuten an, welche Texte darin enthalten sind. So enthält z. B. das Teilkorpus „11–12_1-obd-PV-X“ Texte aus dem 11. und der ersten Hälte des 12. Jahrhunders („11–12_1“), aus der oberdeutschen Sprachregion („obd“), Prosaund Vers-Texte („PV“), aus dem Erweiterungskorpus („X“, d. h. den Texten, die nicht Teil von MiGraKo sind).

Suche nach Pronominaladverbien in Kontakt- vs. Distanzstellung

Pronominaladverbien wie *dâran* ‚daran‘, *dârbî* ‚dabei‘ etc. kommen sowohl in Kontakt- wie auch in Distanzstellung vor, d. h. der pronominale (*dâ*) und der adverbiale (*bî*) Bestandteil stehen entweder adjazent oder sind durch andere Wörter getrennt, s. (6).

Option „Wort endet auf Eingabe“ sucht. Flexionssuffixe werden dadurch ausgeschlossen, da sie im Lemma nicht gesondert markiert sind.

12 Die abfragbaren Metadaten können über das „i“-Symbol eines Treffers eingesehen werden und sind zudem hier aufgelistet: https://www.linguistics.ruhr-uni-bochum.de/rem/documentation/metadata.html.

(6) a. Kontaktstellung mit Zusammenschreibung:
dabi *fol man den irkenn en* (M113y-N1)
‚daran soll man den erkennen'

b. Kontaktstellung mit Getrenntschreibung:
da bî *mach man dich erchennen* (M068-N1)
‚daran kann man dich erkennen'

c. Distanzstellung:
da *erchennet er mich* ***bi*** (M312-G1)
‚daran erkennt er mich'

In ReM sind Pronominaladverbien durch die Form ihrer Lemmata besonders markiert. In Kontaktstellung ist beispielsweise das Lemma des pronominalen Teils von *dârbî* „dâr/+bî", in Distanzstellung „dâr/.+bî" (mit zusätzlichem Punkt); das „/+" zeigt an, dass der zweite Teil noch fehlt zur vollständigen Form. Analog ist das Lemma des adverbialen Teils „bî/+dâr" bzw. „bî/.+dâr".

Möchte man in ANNIS nach Annotationen suchen, so ist die allgemeine Form `Merkmal = "Wert"`. Als Treffer liefert ANNIS dann alle Instanzen, die mit dem entsprechenden Merkmalswert annotiert sind. So kann man z. B. mit den Anfragen in (7) nach Beispiel (6a) suchen.

(7) a. `lemma = "dâr/+bî"`

b. `tok_dipl = "dabi"`

c. `pos = "PAVD"`

Die Anfrage in (7c) erfasst sämtliche Pronominaladverbien in beliebiger Stellung. Möchte man nach sämtlichen Pronominaladverbien, aber in spezifischer Stellung suchen, so muss man die Lemmaformen mit sogenannten „Regulären Ausdrücken" abfragen. Reguläre Ausdrücke erlauben den Einsatz von Wildcards wie „." (ein Punkt für ein einzelnes, beliebiges Zeichen), „`.*`" (Punkt gefolgt von Stern für beliebige Zeichenfolgen, beliebig lang), „`.+`" (beliebige Zeichenfolgen, mindestens 1 Buchstaben lang), „`A?`" (für ein optionales A) oder „`(A|B)`" (für A oder B). Die allgemeine Form für die Suche nach Regulären Ausdrücken ist `Merkmal = /Wert/`.

(8) zeigt einen ersten Versuch, nach allgemeinen Pronominaladverbien in Distanzstellung zu suchen. Statt der konkreten Buchstaben „dâr" und „bî" werden hier die Wildcards „`.*`" eingesetzt.

(8) `lemma = /.*/.+.*/` (vorläufige Version)

Die Anfrage ist allerdings so noch fehlerhaft, da innerhalb eines Regulären Ausdrucks alle speziellen Sonderzeichen markiert werden müssen, wenn nach ihnen gesucht werden soll. In (8) betrifft das die Zeichen „/“, „.“ und „+“ (der Schrägstrich ist auch ein Sonderzeichen, da er den Regulären Ausdruck begrenzt). Um zu markieren, dass diese Zeichen hier keine Wildcards oder Begrenzer sind, setzt man „\“ vor die betreffenden Zeichen, s. (9).

(9) `lemma = /.*\/\.\+.*/` (vorläufige Version)

Schließlich gibt es (in der aktuellen Version von ANNIS) noch die Sonderregel, dass der Schrägstrich hier über seine Unicode-Kodierung in der Form „`\x2F`“ eingegeben werden muss. Die endgültige Form der Suchanfrage ist daher (10). Im Abschnitt 3.3 zeigen wir Ergebnisse dieser Anfrage.

(10) `lemma = /.*\x2F\.\+.*/` (finale Version)

Wenn man wissen möchte, welche der drei Versionen in (7) am häufigsten vorkommt, so muss man die Ebene „tokenization“ in die Anfrage einbeziehen. Pronominaladverbien, die in der originalen Handschrift zusammengeschrieben wurden, haben hier den Wert „MS1“ am ersten Bestandteil und „MS2“ am zweiten, für (moderne) „Multiverbierung durch Spatium“. Ein Ausdruck der Form `A _=_ B` besagt, dass ein Element sowohl mit dem Merkmal A als auch mit B annotiert sein soll. Damit kann man die beiden Bedingungen – die Form des Lemmas und den tokenization-Wert – zusammenführen, s. (11a). Da nicht abfragbar ist, dass ein Element *kein* Merkmal auf der Ebene „tokenization“ hat, ziehen wir die Trefferzahl der Zusammenschreibungen von der Gesamttrefferzahl ab. Beispiel (11) zeigt die Anfragen und Frequenz-Ergebnisse für alle drei Versionen aus (7).

(11) a. Kontaktstellung mit Zusammenschreibung: 3.702 Treffer
`lemma = /.*\x2F\+.*/ _=_ tokenization = "MS1"`

b. Kontaktstellung mit Getrenntschreibung: 22.654–3.702 = 18.952 Treffer
`lemma = /.*\x2F\+.*/`

c. Distanzstellung: 1980 Treffer
`lemma = /.*\x2F\.\+.*/`

Prä- vs. Postpositionen

Bisher waren die Suchen beschränkt auf einzelne Wörter. Für komplexere linguistische Phänomene gibt es Such-Operatoren, mit denen Beziehungen

zwischen mehreren Wörtern ausgedrückt werden können. Solche Operatoren benötigt man, wenn man sich z. B. für die Entwicklung der Adposition interessiert, d. h. für ihre Position vor vs. nach ihrem Bezugswort (Prä- vs. Postpositionen). Adpositionen sind in ReM generell mit der Wortart „APPR" annotiert, unabhängig von ihrer Position. Die Anfrage in (12) findet daher sämtliche Adpositionen im Korpus (149.728 Treffer).

(12) `pos = "APPR"`

Um die beiden Positionen zu unterscheiden, müssen Heuristiken angewendet werden. Wir beginnen mit der Suche nach Präpositionen. Dafür muss einerseits ausgedrückt werden, dass nach der Präposition nominale Elemente folgen müssen (das Bezugswort der Präposition). In erster Annäherung definieren wir als nominale Elemente:

- Nomen (mit der Wortart „NA"), Eigennamen („NN")
- substituierende Pronomen (z. B. „DDS", „DGS", etc.: alle mit „D" beginnend und „S" endend, dazu Personalpronomen („PPER"), Reflexivpronomen („PRF") u. a.)
- Artikelwörter, die eine Nominalphrase einleiten (z. B. „DDA", „DDART" etc.)
- attributive Adjektive („ADJA")

Diese Elemente erfassen wir über einen komplexen Regulären Ausdruck mit einer Disjunktion über alle Alternativen, s. (13).

(13) `pos=/(N.|PPER|PRF|PG|PI|PW|D.*|ADJA)/`

Jetzt müssen die beiden Teile noch zusammengefügt werden. Das passiert mit Hilfe des Präzedenz-Operators „`.`" (einem Punkt): der Ausdruck `A . B` besagt, dass ein Element, das die Bedingung A erfüllt, direkt vor einem Element, das B erfüllt, steht, s. (14).[13]

(14)
```
pos = "APPR".
pos = /(N.|PPER|PRF|PG|PI|PW|D.*|ADJA)/
```

Um sicher zu sein, dass sich die Adposition nicht auf ein vorhergehendes Element bezieht und nur zufällig ein nominales Element nachfolgt, schließen wir zusätzlich nominale Elemente direkt vor der Adposition aus. Dazu nutzen wir

13 In (14) sind die beiden Bedingungen der Lesbarkeit wegen auf zwei Zeilen verteilt. Das hat keinen Einfluss auf die Anfrage.

die Negation „!“ (ein Ausrufezeichen), die besagt, dass ein Merkmal *nicht* den entsprechenden Wert haben darf, s. (15).

(15)
```
pos != /(N.|PPER|PRF|PG|PI|PW|D.*|ADJA)/ .
pos = "APPR".
pos = /(N.|PPER|PRF|PG|PI|PW|D.*|ADJA)/
```

Mit diesem Ausdruck erhalten wir 86.638 Treffer, beispielsweise den in (16). In der ersten Zeile mit dem Beleg sind die drei Wörter, die den Suchausdruck erfüllen, fett gedruckt, in der zweiten Zeile ist für alle Wörter des Belegs die jeweilige Wortart angegeben.

(16) *der* ***ift*** ***ane*** ***zwiuel*** *îmer ſalich* (M010-N1)
DDS VAFIN APPR NA AVD ADJD
‚der ist ohne Zweifel immer selig‘

Die Suche nach Postpositionen verläuft analog, es müssen nur die erste und die dritte Zeile von (15) vertauscht werden. Wir erhalten dann 1.195 Treffer, darunter den in (17).

(17) *vnd díe gemeýnde der* ***schulde halb numm’*** (M533-N0)
KON DDART NA DDART NA APPR AVD
vírdorbín were
VVPP VAFIN
‚und die Gemeinde der Schuld halber niemals verdorben wäre‘

Eine nähere Inspektion der Treffer zeigt allerdings, dass quasi alle Treffer ungewollt sind und tatsächlich Instanzen von Präpositionen darstellen.

Da die Stellung von Adpositionen in ReM nicht explizit annotiert ist und wir daher heuristisch vorgehen mussten, war zu erwarten, dass die Suchanfrage einerseits ungewollte Treffer („false positives“) erzielt und andererseits Instanzen verpasst werden („false negatives“). (18) zeigt Beispiele für beide Fälle. Viele der verpassten Instanzen von Präpositionen involvieren Adverbien, aber natürlich schließen wir mit unserer Anfrage auch alle Instanzen aus, in denen die Präposition auf ein vorhergehendes nominales Element folgt.

(18) a. *Nu intheizeſ tu unſ* neheina ***ſicherheit uone*** (M182A-N1)
AVD VVFIN PPER PPER DIA NA APPR
danne
AVD
‚nun versprichst du uns keine Sicherheit von da‘
Präposition, false negative

b. *mit* ***worten mit*** · *werchen oder mit boesen gedanchen* (M096-N1)
APPR NA APPR $_ NA KON APPR ADJA NA
‚mit Worten, mit Werken oder mit bösen Gedanken'
Postposition, false positive

Diachrone Verläufe
Besonders interessant ist es, sich diachrone Verläufe anzuschauen. Dazu muss auf Metadaten Bezug genommen werden. Als Beispiel soll die Abfolge Vollverb-Auxiliar dienen. Während im heutigen Deutsch die unmarkierte Wortfolge im Nebensatz (meist) Vollverb > Auxiliar ist, kommt im Mittelhochdeutschen auch die umgekehrte Abfolge häufig vor. Um einen Eindruck von der diachronen Entwicklung zu bekommen, teilen wir das Korpus in Blöcke à 50 Jahre auf. Von insgesamt 2,5 Mio. Tokens im Korpus stammen 1,9 Mio. Tokens aus Texten, die bis auf 50 Jahre genau zeitlich zugeordnet sind. Die entsprechenden Texte sind z. B. annotiert mit „date = 14,1" (erste Hälfte 14. Jh., also 1300–1350). Die restlichen Tokens sind auf 100 Jahre genau klassifiziert bzw. 10 Texte haben aktuell keine Information zur zeitlichen Einordnung. Wir beschränken unsere Suche auf die auf 50 Jahre datierten Texte.

Da pro Zeitfenster unterschiedlich viele Instanzen von Vollverb-Auxiliar-Kombinationen vorkommen, ermitteln wir zunächst unabhängig von der Reihenfolge die Gesamtzahl an Vorkommen in einem Zeitfenster. Dazu wird der Operator „^" verwendet: `A ^ B` bedeutet, dass entweder `A B` vorangeht oder umgekehrt. Bedingungen über Metadaten werden mit dem Operator „&" angefügt, außerdem wird dem Merkmalsnamen `meta::` vorangestellt, s. (19). (Die Anfrage ist wieder nur eine grobe Annäherung und erfasst beispielsweise auch Verbsequenzen in Hauptsätzen.)

(19)
```
pos = "VAFIN" ^ pos = /VV.*/ &
meta::time = "11,2"
```

Diese Anfrage ergibt 149 Treffer in 5 Texten (der abgefragte Zeitraum enthält insgesamt 25.758 Tokens in 8 Texten). Als nächstes schränken wir die Suche auf eine der beiden Abfolgen ein (20).

(20)
```
pos = "VAFIN" . pos = /VV.*/ &
meta::time = "11,2"
```

Diese Anfrage hat 52 Treffer in 4 Texten. Damit überwiegt im frühen Mittelhochdeutsch die Auxiliar-finale Stellung mit zwei Dritteln der Vorkommen deutlich.

Man könnte nur für alle 50-Jahre-Zeitfenster entsprechende Anfragen machen und die Anzahl der Treffer notieren. Praktischerweise bietet aber ANNIS eine Frequenz-Analyse, die uns weite Teile der Arbeit abnimmt. Diese Analyse wird im Abschnitt 3.3 vorgestellt.

Zum Abschluss werfen wir nochmal einen Blick auf die *minne*-Anfrage in (5). Die Anfrage der vereinfachten Suchmaske wurde übersetzt in die entsprechende ANNIS-Anfrage, dabei werden Reguläre Ausdrücke eingesetzt wie z. B. `.*12.*`, so dass die Angabe „12. Jahrhundert" den Wert „12" aber auch „12,1" und „12,2–13,1" erfasst.

3.2 Export der Suchergebnisse

Oft möchte man die Ergebnisse einer Suchanfrage weiter verwenden, z. B. als Belege in einer wissenschaftlichen Arbeit oder um die Ergebnisse außerhalb des Suchtools mit weiteren Informationen zu annotieren, z. B. ob es sich um „true positives" (korrekte Treffer) oder um „false positives" (ungewollte Treffer) handelt.

ANNIS bietet verschiedene Export-Optionen an. Dazu formuliert man eine Anfrage, klickt dann unter dem Suchfenster auf „More" und wählt „Export" aus. Im neuen Reiter „Export" im Fenster rechts können dann verschiedene Optionen ausgewählt werden. Im Folgenden stellen wir einen der Exporter vor.

Text-Export

Wir beginnen mit der Beispiel-Anfrage `lemma = "hûs"`. Als Exporter bietet sich hier der „GridExporter" an.[14] Unter „Annotation Keys" trägt man die Annotationsebene(n) ein, die man exportieren möchte, in unserem Fall „tok_anno", dazu unter „Parameters" eintragen: „numbers=false". Klickt man auf „Perform Export", wird die entsprechende Suche durchgeführt. Anschließend kann die Datei mit den Daten per Klick auf „Download" abgespeichert werden.

Die Datei nummeriert (mit „0" beginnend) und listet alle Treffer mit einem voreingestellten Kontext von fünf Tokens links und rechts des Suchausdrucks (s. Abb. 5.6). Der Kontext lässt sich bis auf 20 Tokens erweitern.

14 Der „SimpleTextExporter" kann mit ReM nicht verwendet werden, da ReM für den Text nicht die „normale" ANNIS-Textebene nutzt, sondern eigene Ebenen „tok_dipl" und „tok_anno" definiert.

```
0.   tok_anno   irkihtost den toten in dem hus . Herro irkike mih fon
1.   tok_anno   getriu . noh nehein gotes hus . den gotes lichinamen .
2.   tok_anno   ze heili . die din hus uuerden sculen . Ne simul
```

Abb. 5.6: Text-Export der Anfrage `lemma = "hûs"`.

Export von Annotationen
Falls auch Annotationen exportiert werden sollen, trägt man beim „GridExporter" unter „Annotation Keys" die entsprechenden Ebenen durch Komma getrennt ein, z. B. „tok_anno,inflection". Zusätzlich kann man unter „Parameters" Metadaten angeben, z. B. „metakeys=time,language-area,text".[15] Abbildung 5.7 zeigt einen Ausschnitt der Export-Datei (mit dem Kontext auf zwei Tokens gesetzt).[16]

```
0. tok_anno     in[1-1] dem[2-2]          hus [3-3]    .[4-4]  Herro[5-5]
   inflection   --[1-1] Neut.Dat.Sg[2-2] Dat.Sg[3-3]          Nom.Sg[5-5]
   meta::time   12,1
   meta::language-area alemannisch
   meta::text   Rheinauer Gebete

1. tok_anno     nehein[1-1]         gotes[2-2]  hus[3-3]    .[4-4] den[5-5]
   inflection   Neut.Akk.Sg.0[1-1] Gen.Sg[2-2] Akk.Sg[3-3]        Masc.Akk..
   meta::time    11,2-12,1
   meta::language-area bairisch
   meta::text   Wessobrunner Glaube u. Beichte I
```

Abb. 5.7: Annotations-Export der Anfrage `lemma = "hûs"`.

3.3 Frequenz-Analyse

Wie erwähnt, bietet ANNIS eine einfache statistische Analyse. Dazu stellt man zunächst eine Anfrage, klickt dann unter dem Suchfenster auf „More" und wählt

15 Dabei sollte „numbers" nicht auf „false" gesetzt werden, sonst ist eine korrekte Zuordnung der Annotationen zu den Tokens bei leeren Annotationsfeldern nicht möglich.
16 Die visuelle Alignierung der Tokens und ihrer Annotation ist manuell gemacht. ANNIS liefert mit den Zahlen „[1–1]" etc. die dazu notwendige Information.

„Frequency Analysis“ aus. Im Folgenden zeigen wir für einige der Suchanfragen aus Abschnitt 3.1, wie interessante Frequenz-Analysen erstellt werden.

hûs: Wortformen eines Lemmas

Als erste Beispielanfrage nehmen wir `lemma = "hûs"`. Nach Klick auf „Frequency Analysis“ öffnet sich im Hauptfenster rechts ein neuer Reiter. Im oberen Bereich stehen Informationen zur Anfrage: die ausgewählten Korpora sowie die Anfrage selbst.

Download as CSV

84 items with a total sum of 1348 (query on 11-12_1-obd-PV-X, 11-12_1-rhfrhess-PV-X, 11_2-12_1-obd-F

rank	#1|tok_anno	count
1	hus	429
2	huse	166
3	huos	134
4	hvs	102
5	huses	76
6	hvse	66
7	haus	32
8	hvses	24
9	Haus	22
10	huis	22

Abb. 5.8: Frequenz-Analyse der Wortformen des Lemmas *hûs*.

Im unteren Bereich können die Merkmale angegeben werden, für deren Frequenz-Analyse man sich interessiert. Automatisch vorgegeben sind Merkmale, die in der Anfrage selbst vorkommen, hier also „lemma“. Klickt man unten auf „Perform frequency analysis“, so erhält man ein Balkendiagramm, das die Verteilung der verschiedenen Merkmalswerte anzeigt. Da wir nur nach genau einem Lemma gesucht haben, ist eine Frequenz-Analyse hier uninteressant und es gibt nur genau einen Balken. Klicken wir daher auf „New Analysis“, löschen das vorgegebene Merkmal (dazu den Eintrag anklicken und unten auf „Delete selected row(s)“ klicken) und fügen stattdessen ein anderes Merkmal hinzu, z. B. die konkrete Wortform, „tok_anno“. Dazu unten auf „Add“ klicken und in der Spalte „Selected annotation of node“ „tok_anno“ eintragen. Das Balkendiagramm zeigt nun für das Lemma *hûs* alle vorkommenden Wortformen mit ihrer Frequenz. Spitzenreiter ist die Form *hus* mit 429 Vorkommen, gefolgt von *huse* (166 Vorkommen). Auf Platz 3 steht eine Wortform, deren Schreibung eine Diphthongierung bereits andeutet: *huos* (134 Vorkommen) (s. Abb. 5.8). Die Rohdaten der Analyse können im csv-Format heruntergeladen und extern weiterverarbeitet werden.

Pronominaladverbien: Instanzen eines Regulären Ausdrucks

Als weiteres Beispiel dient die Anfrage (10) aus Abschnitt 3.1.2: `lemma = /.*\x2F\.\+.*/`. Der Reguläre Ausdruck sucht nach Pronominaladverbien in Distanzstellung und deckt eine Vielzahl an Lemmata ab, für die wir uns nun interessieren. Die Frequenz-Analyse für das Merkmal „lemma“ zeigt, dass die häufigste Form *dârinne* mit 371 Treffern ist, gefolgt von *dârane* (222 Treffer) und *dârmit(e)* (219 Treffer).

Auxiliar/Vollverb: Verteilungen von Wortart-Kombinationen

Zum Abschluss betrachten wir noch die Anfragen zur Abfolge von Auxiliar und Vollverb, vgl. (20), nun aber ohne Einschränkung des Zeitraums: `pos = "VAFIN" . pos = /VV.*/`. Die Frequenz-Analyse bietet zunächst zwei automatisch generierte Einträge nach den Merkmalen von „pos“, nummeriert gemäß der Reihenfolge in der Anfrage. Der Wert des ersten Merkmals ist fix („VAFIN“), aber das zweite Merkmal umfasst alle Wortarten, die mit „VV“ beginnen. Tabelle 5.3 zeigt die Verteilung: die überwiegende Mehrheit der Instanzen stellt die Kombination „VAFIN VVPP“ dar. Die Tabelle zeigt außerdem die Verteilung der umgekehrten Reihenfolge, auch hier deckt „VVPP VAFIN“ den Großteil der Fälle ab.

Die große Gruppe der Kombination mit VVPP kann man sich nun noch genauer anschauen, indem man beispielsweise das Lemma des Auxiliars mit einbezieht. Dazu muss man in der Frequenz-Analyse einen weiteren Eintrag

hinzufügen (Klick auf „Add“), in der neuen Zeile trägt man die Knotennummer („1“ im Falle der Anfrage „VAFIN VVPP“, „2“ für „VVPP VAFIN“) und als Annotation „lemma“ ein. Tabelle 5.4 zeigt hier deutliche Unterschiede im Gebrauch von *sîn* und *wërden*.

Tab. 5.3: Abfolge von Auxiliar und Vollverb in verschiedenen Kombinationen.

Kombination	Frequenz	Kombination	Frequenz
VAFIN VVPP	9.106	VVPP VAFIN	10.517
VAFIN VVFIN	158	VVFIN VAFIN	200
VAFIN VVPS	109	VVPP VAFIN	207
VAFIN VVINF	106	VVINF VAFIN	136
VAFIN VVIMP	64	VVIMP VAFIN	11

Tab. 5.4: Lemmata des Auxiliars in Kombinationen von Auxiliar und Partizip (VVPP).

Lemma	Auxiliar > VVPP	VVPP > Auxiliar
haben	3.401 (37 %)	3.925 (37 %)
sîn	1.907 (21 %)	2.978 (28 %)
wërden	2.454 (27 %)	2.353 (22 %)
wësen	1.344 (15 %)	1.258 (12 %)

Abb. 5.9: Diachroner Verlauf der Kombination von Auxiliar und Vollverb; die absoluten Frequenzen sind in die Balken eingeschrieben.

Schließlich kann man als weiteren Faktor die zeitliche Einordnung hinzufügen, um den diachronen Verlauf zu untersuchen. Dazu gegebenenfalls alle Einträge löschen und stattdessen unten unter „Metadata" in das leere Feld „time" eintragen (oder auf „Select" klicken und aus der Liste auswählen). Abbildung 5.9 zeigt den diachronen Verlauf in Form eines Säulendiagramms (nur die genaueren 50-Jahre-Angaben sind berücksichtigt). Während zu Beginn ein Trend zu mehr initialem Auxiliar vorzuliegen scheint, liegt gegen Ende die verb-initiale Stellung wieder deutlich vorne. Allerdings ist zu beachten, dass die Konstruktionen nur in den mittleren Bereichen einigermaßen häufig belegt sind.

4 Fazit und Ausblick

Die in den letzten Jahren entstandenen und noch im Entstehen befindlichen historischen Referenzkorpora des Deutschen bieten eine hinlänglich umfangreiche, verlässliche und handschriftengetreue Datenbasis deutschsprachiger Texte älterer Sprachstufen. Damit erlauben sie historiolinguistische und mediävistische Recherchen in einem Maße, das weit über das hinausgeht, was bisher möglich war. Unabhängig von der großen Zahl der Wortformen bietet auch die Annotationstiefe bisher nicht dagewesene Recherchemöglichkeiten auf vielen grammatischen Ebenen. Die Abfrage mit Hilfe von Korpustools wie ANNIS erlaubt zudem quantitative Untersuchungen auf großen Datenmengen.

Die Verfügbarkeit von Referenzkorpora bietet zudem den großen Vorteil, Ergebnisse reproduzierbar und vergleichbar zu machen. Wünschenswert wäre hier, dass Benutzer die Ergebnisse eigener, externer Untersuchungen der Korpusdaten in die Korpora zurückspeisen könnten, so dass auch diese Information für andere nutzbar und nachprüfbar wird.

Ebenso wäre es wünschenswert, wenn Benutzer mit ihren eigenen Korpora, sofern sie den Maßstäben der Referenzkorpusprojekte genügen, die Referenzkorpora ergänzen könnten.

Literatur

Bollmann, Marcel, Florian Petran, Stefanie Dipper & Julia Krasselt (2014): CorA: A webbased annotation tool for historical and other non-standard language data. In Kalliopi Zervanou, Cristina Vertan, Antal van den Bosch & Caroline Sporleder (Hrsg.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 86–90. Gothenburg, Sweden: Association for Computational Linguistics.

Brants, Thorsten & Wojciech Skut (1998): Automation of treebank annotation. In David M.W. Powers (Hrsg.), *Proceedings of the Joint Conference on New Methods in Natural Language Processing and Computational Language Learning, NeMLaP3/CoNLL98*, 49–57. Sydney, Australia: Association for Computational Linguistics.

Brugman, Hennie & Albert Russel (2004): Annotating multimedia/multi-modal resources with ELAN. In Maria Theresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Cost & Raquel Silva (Hrsg.), *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, 2065–2068. Paris: European Language Resources Association (ELRA).

Dipper, Stefanie, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller & Klaus-Peter Wegera (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal of Language Technology an Computational Linguistics* 28(1), 85–137.

Donhauser, Karin (2015): Das Referenzkorpus Altdeutsch: Das Konzept, die Realisierung und die neuen Möglichkeiten. In Jost Gippert & Ralf Gehrke (Hrsg.), *Historical corpora. Challenges and perspectives*, 35–49. Tübingen: Narr.

Grimm, Jacob & Wilhelm Grimm (1854–1971): *Deutsches Wörterbuch. 16 Bde. in 32 Teilbänden und Quellenverzeichnis*. Leipzig. [Nachdruck, DTV: München 1984]. Online: woerterbuchnetz.de/DWB.

Herbers, Birgit (2016): „Referenzkorpus Deutsche Inschriften" – Chancen und Grenzen der Auswertung. In Sarah Kwekkeboom & Sandra Waldenberger (Hrsg.), *Perspektiv-Wechsel oder: Die Wiederentdeckung der Philologie, Band 1: Sprachdaten und Grundlagenforschung in der Historischen Linguistik*, 27–41. Berlin: ESV.

Klein, Thomas & Stefanie Dipper (2016): Handbuch zum Referenzkorpus Mittelhochdeutsch. *Bochumer Linguistische Arbeitsberichte* 19.

Krause, Thomas & Amir Zeldes (2016): ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31, 118–139. http://dsh.oxfordjournals.org/content/31/1/118.

Lexer, Matthias (1872–1878): *Mittelhochdeutsches Handwörterbuch*. 3 Bde. Leipzig. [Nachdruck, Hirzel: Stuttgart 1992]. Online: woerterbuchnetz.de/lexer.

Peters, Robert & Norbert Nagel (2014): Das digitale ‚Referenzkorpus Mittelniederdeutsch/Niederrheinisch (ReN)'. In Vilmos Ágel & Andreas Gardt (Hrsg.), *Paradigmen der Sprachgeschichtsschreibung*, 165–175. Berlin, Boston: de Gruyter.

Petran, Florian, Marcel Bollmann, Stefanie Dipper & Thomas Klein (2016): ReM: A reference corpus of Middle High German – corpus compilation, annotation, and access. *Journal of Language Technology an Computational Linguistics* 31(2), 1–15.

Schmidt, Thomas (2001): The transcription system EXMARaLDA: An application of the annotation graph formalism as the basis of a database of multilingual spoken discourse. In Stephen Bird et al. (Hrsg.), *Proceedings of the IRCS Workshop on Linguistic Databases*, 219–227. Philadelphia, PA: Institute for Research in Cognitive Science, University of Pennsylvania.

Schröder, Ingrid (2014): Das Referenzkorpus: Neue Perspektiven für die mittelniederdeutsche Grammatikographie. In Vilmos Ágel & Andreas Gardt (Hrsg.), *Paradigmen der Sprachgeschichtsschreibung*, 150–164. Berlin, Boston: de Gruyter.

Solms, Hans-Joachim & Klaus-Peter Wegera (1998): Das Bonner FrühneuhochdeutschKorpus. Rückblick und Perspektiven. In Rolf Bergmann (Hrsg.), *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung, 1. und 2. November 1996*, 22–39. Stuttgart: S. Hirzel.

von Steinmeyer, Elias (Hrsg.) (1916): *Die kleineren althochdeutschen Sprachdenkmäler*. Berlin: Weidmann. Nachdruck Dublin/Zürich: Weidmann, 1971.

Wadstein, Elis (Hrsg.) (1899): *Kleinere altsächsische sprachdenkmäler. mit anmerkungen und glossar*. Norden/Leipzig: Soltau.

Wegera, Klaus-Peter (2000): Grundlagenprobleme einer mittelhochdeutschen Grammatik. In Werner Besch, Anne Betten, Oskar Reichmann & Stefan Sonderegger (Hrsg.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* 2. Teilband (Handbücher zur Sprach- und Kommunikationswissenschaft 2.2), 1304–1320. Berlin, New York: de Gruyter.

Roland Kehrein und Lars Vorberger

6 Dialekt- und Variationskorpora

Abstract: Dialekt- und Variationskorpora enthalten Sprachdaten zu bestimmten Varietäten des Deutschen und/oder bilden sprachliche Variation in der arealhorizontalen oder der situativ-vertikalen Dimension ab. Somit stellen sie für die (moderne) Regionalsprachenforschung eine wichtige Datengrundlage dar. Im vorliegenden Beitrag erfolgt nach einer kurzen begrifflichen Erläuterung von Dialekt- und Variationskorpora sowie deren Spezifika hinsichtlich ihrer Zusammenstellung (Datenerhebung) eine schlaglichtartige chronologische und systematische Präsentation von Dialektatlaskorpora, variationslinguistischen Tonkorpora und modernen variationslinguistischen Korpora. Über deren Inhalt hinaus werden jeweils auch die Formen ihrer Erschließung sowie ihre Zugänglichkeit im Rahmen von Online-Plattformen beschrieben.

Keywords: Sprachvariation, Dialekt, Dialektatlas, Deutsch heute, Pfeffer, SiN, REDE, Regionalsprache, Sprachatlas, Wenker, Zwirner

1 Einleitung

Im vorliegenden Beitrag wird es um Dialekt- und Variationskorpora gehen, Korpora von Sprachdaten also, in denen sprachliche Variation und/oder sprachliche Varietäten erfasst sind. Die Frage, was ein Korpus ist, wird im vorliegenden Band ausführlich behandelt. Aus diesem Grund werden wir uns nicht intensiver mit dem Korpusbegriff auseinandersetzen. Wir legen eine Minimaldefinition zugrunde, nach der ein Korpus die folgenden Eigenschaften hat:

- Es handelt sich um eine „geschlossen verfügbare Sammlung von Texten" (Lehmann 2007: 17).
- Diese Texte gehören einer bestimmten Kategorie an.
- Es kann sich um geschriebene Sprache und um gesprochene Sprache handeln, wobei letztere in irgendeiner Form transkribiert sein muss (vgl. Friginal & Hardy 2014: 20).

Roland Kehrein, Forschungszentrum Deutscher Sprachatlas, Pilgrimstein 16, D-35032 Marburg, E-Mail: kehrein@uni-marburg.de
Lars Vorberger, Forschungszentrum Deutscher Sprachatlas, Pilgrimstein 16, D-35032 Marburg, E-Mail: lars.vorberger@deutscher-sprachatlas.de

https://doi.org/10.1515/9783110538649-007

Korpora sind in vielen Fällen zusätzlich über Metadaten sowie durch Annotationen erschlossen und können dadurch mit Recherchewerkzeugen durchsucht werden. Da diese Eigenschaften aber nicht auf alle in diesem Beitrag präsentierten Korpora zutreffen, schließen wir sie nicht in unsere Definition mit ein.

Was versteht man auf der anderen Seite nun unter Variation von Sprache? Diese lässt sich auf mehreren Dimensionen beobachten und beschreiben. Unterschieden werden (vgl. Coseriu 1992: 280):

- Sprachvariation in Abhängigkeit vom geographischen Raum: diatopische Dimension,
- Sprachvariation in Abhängigkeit von der Zeit: diachrone Dimension,
- Sprachvariation in Abhängigkeit von der Kommunikationssituation: diaphasische Dimension und
- Sprachvariation in Abhängigkeit von sozialen Faktoren wie Alter/Generation, Bildung, Geschlecht usw.: diastratische Dimension.

Die vor allem in jüngster Zeit für das Deutsche intensiv untersuchte Sprachvariation auf der vertikalen Dimension zwischen Standardsprache und Dialekt (vgl. etwa Lenz 2003; Lameli 2004; Christen et al. 2010; Kehrein 2012; Rocholl 2015; Vorberger 2017, um nur die rezent vorgelegten Buchpublikationen zu nennen) integriert Aspekte der diaphasischen und diastratischen Dimension, da sie von situativen Faktoren abhängt, zu denen beispielsweise auch Eigenschaften des Gesprächspartners gehören. Dialekt- und Variationskorpora sind daher entsprechend bestimmbar als abgeschlossene Sammlungen von Texten gesprochener Sprache, die zu Dokumentationszwecken oder für bestimmte Forschungsziele (z. B. Erstellung von Sprachatlanten, Untersuchung intersituativer Sprachvariation) erhoben wurden. In den meisten Fällen handelt es sich nicht um Zusammenstellungen vorhandener Texte, sondern um eigens zu den genannten Zwecken angefertigte Neuerhebungen.[1] Die wichtigste Variationsdimension im Deutschen – dies ist aus der Geschichte des Gesamtsprachsystems Deutsch zu erklären – ist die diatopische Dimension (vgl. auch Bellmann 1994: 2). Dialekt- und Variationskorpora enthalten daher immer Sprachdaten aus verschiedenen Regionen und Orten eines definierten Untersuchungsgebietes des deutschen Sprachraums. Um die Vergleichbarkeit der Daten zu gewährleisten, werden bei

---

1 Ausnahmen bilden Sammlungen von Sprachaufzeichnungen, die ursprünglich nicht für sprachwissenschaftliche Zwecke angefertigt wurden. Als Beispiele lassen sich Zusammenstellungen (i) von Mitschnitten von Gemeinderatssitzungen (vgl. Lameli 2004) sowie (ii) von Notrufgesprächen (vgl. Kehrein 2006; Christen, Hove & Petkova 2015) nennen, die dann jeweils hinsichtlich bestimmter Fragestellungen ausgewertet wurden. Diese Korpora sind aus verschiedenen Gründen in der Regel aber nicht zugänglich, weshalb sie im vorliegenden Beitrag keine weitere Berücksichtigung finden.

der Datensammlung in der Regel klare Kriterien an die Informantenauswahl und an die Erhebungsmethoden angelegt (vgl. auch Hunston 2008). Bei der Informantenauswahl wird zunächst einmal ein besonderes Augenmerk auf die Orts-/Regionsfestigkeit gelegt. Das bedeutet, dass die Informanten sowie möglichst auch ihre Eltern und Großeltern aus dem Ort/der Region stammen und sie keine längeren Aufenthalte in anderen Regionen gehabt haben sollten. Darüber hinaus werden Informanten beispielsweise nach ihrem Alter bzw. ihrer Generationenzugehörigkeit, ihrem (früheren) Beruf und/oder ihrem Bildungsniveau ausgewählt. Die außersprachliche Variable Geschlecht wird bei großräumig angelegten Projekten in der Regel nicht systematisch variiert. Dies hat zum einen forschungspraktische Gründe, denn die Informantenzahl würde sich jeweils verdoppeln. Zum anderen ist noch nicht abschließend geklärt, inwieweit das Geschlecht einen Einfluss auf das variative Sprachverhalten und die Dialektkompetenz hat.[2]

Was die Erhebungsmethoden angeht, so lässt sich hinsichtlich Dialekt- und Variationskorpora auf einer übergeordneten Ebene zunächst einmal die direkte von der indirekten Erhebung unterscheiden. Während bei ersterer die Sprachdaten (meist vor Ort) im unmittelbaren Kontakt mit dem Informanten erfasst werden (Sprachaufzeichnung oder direkte Transkription), arbeitet man bei der zweiten Methode mit Fragebogen, die von den Informanten schriftlich auszufüllen sind. Über diese Möglichkeiten der Datenerfassung hinaus lässt sich differenzieren, welcher Bereich sprachlicher Variation erhoben werden soll. Traditionell ging es um die Dialekte, und zwar um eine möglichst „bodenständige“ Form der Dialekte. Es wurde die Dialektkompetenz der Informanten in direkten oder indirekten Befragungen erhoben. Solche Befragungen können prinzipiell alle linguistischen Systemebenen zum Gegenstand haben. Sie basieren auf einem standardisierten Set an Abfragewörtern und -sätzen, sodass für alle berücksichtigten Orte und Sprecher vergleichbares Sprachmaterial vorliegt. In jüngeren Projekten werden per Abfrage und durch lautes Vorlesen auch Daten zur Standardkompetenz der Informanten, ihr individuell bestes Hochdeutsch, erhoben.[3] Außer der Kompetenzabfrage können Dialekt- und

---

2 Zu dem Ergebnis, dass der Faktor Geschlecht keine Rolle spielt, gelangt beispielsweise Auer (1990) in seiner Untersuchung zur Sprachvariation in der Konstanzer Stadtsprache. Während im Falle von Dialektatlanten vielfach angenommen wird, der geeignetste und typische Informant sei „non mobile, old, rural, male“ (NORM; vgl. Chambers & Trudgill 1998), geben beispielsweise Jaberg & Jud (1928) an, dass Frauen über eine besonders ausgeprägte und differenziertere Dialektkenntnis verfügten (vgl. Jaberg & Jud 1928: 189–193).

3 Als Vorbild für die Erhebung dieses Gegenstands kann Wilhelm Viëtor mit seiner Reihe „Beiträge zur Statistik der Aussprache des Schriftdeutschen“ (vgl. z. B. Viëtor 1888) gelten. In jüngerer Zeit hat Werner König sowohl mit Informanten des *Sprachatlas von Bayerisch-Schwaben*

Variationskorpora auch Texte frei formulierter Monologe oder kurze (Nach-)Erzählungen enthalten (z. B. die von Fleischer & Gadmer 2002 herausgegebenen Schweizer Aufnahmen des Züricher Phonogrammarchivs). Auch in diesen Fällen zielt die Erhebung meist auf den Dialekt der Informanten. In jüngeren Projekten schließlich werden noch Daten zum variativen Sprachgebrauch der Informanten erhoben. Aus Gründen der Vergleichbarkeit handelt es sich um „freie" Gespräche, die in Situationen aufgezeichnet werden, bei denen möglichst viele außersprachliche Variablen konstant gehalten werden können. Dazu gehört vor allem das (sprachbiographische) Interview, in dem ein Explorator als teilnehmender Beobachter die Gesprächsbeiträge der Informanten durch möglichst offene Fragen initiiert, dazu gehören aber auch Situationen, in denen zwei oder mehr Informanten in einer bestimmten Weise miteinander interagieren (z. B. sog. Freundesgespräche oder auch Wegbeschreibungen in der sog. Map Task).

Wir werden im Folgenden Dialekt- und Variationskorpora des Deutschen vorstellen. Dabei werden wir analog zur Chronologie ihrer jeweils ersten Erhebung mit Datensammlungen für Dialekt- bzw. Sprachatlanten beginnen, dann Tonkorpora zu deutschen Varietäten beschreiben oder kurz tabellarisch erfassen (s. Tab. 6.1), bevor drei jüngere Projekte vorgestellt werden, in denen mehrere Variationsdimensionen berücksichtigt werden. Nach der Präsentation der Korpora wird es um ihre Aufbereitung und, was eng damit verknüpft ist, um ihre Erschließung (z. B. über Register) und Zugänglichkeit (z. B. über Datenbank-Frontends) gehen. Exemplarisch werden wir hier die *Datenbank für gesprochenes Deutsch* (DGD2) und die Internetplattform *Regionalsprache.de* (REDE) vorstellen, auf denen jeweils in verschiedenen Korpora recherchiert werden kann.

2 Dialektatlaskorpora

Über die Datensammlungen für Dialektatlanten wurden bereits lange bevor man überhaupt an Korpuslinguistik gedacht hat, Korpora als abgeschlossene Sammlungen von Texten gesprochener Sprache zu Dokumentationszwecken und für bestimmte Forschungsziele erstellt. Allerdings waren diese Datensammlungen in Sprachatlasprojekten nicht als Korpuserstellung gedacht. Daher sind sie in Form der Rohdaten in der Regel auch nicht zugänglich und für weitere Analysen verwendbar. Die einzige, planmäßige (!) Aufbereitung dieser

---

(König 1996–2009) (vgl. dazu Wecker-Kleiner 2009) als auch für den *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland* (König 1989) solche standardorientierten Sprachdaten erhoben. Berücksichtigung findet das individuell beste Hochdeutsch auch in den in Abschnitt 4 beschriebenen Projekten zur Sprachvariation auf der vertikalen Dimension.

Korpora besteht in der Sprachkartenerstellung für den jeweiligen Atlas. Ein häufig nicht unbeträchtlicher Teil der gesammelten Daten bleibt somit unausgewertet (vgl. zu dieser Problematik auch Kunst & Barbiers 2010: 401–402).

2.1 Sprachatlas des Deutschen Reichs

Mit der von Georg Wenker im Zeitraum zwischen 1876 und 1887 durchgeführten Sprachdatenerhebung für seinen *Sprachatlas des Deutschen Reichs* (Wenker 1889–1923) liegt das bis heute umfangreichste Korpus zu den Dialekten einer Einzelsprache vor. Für ihn stand dabei nicht ein dokumentarisches Interesse im Vordergrund, sondern ein wissenschaftliches. Entsprechend bildete die Datensammlung für ihn die Grundlage für die Erstellung und die Auswertung von Sprachkarten. Wenkers erklärtes Ziel war es, eine über große Distanzen hinweg vergleichende Dialektkunde zu betreiben (vgl. Wenker 1886: 190). Dabei war ihm vollkommen klar, dass eine detaillierte Totalerhebung aller Dialekte unmöglich war: Er beschränkte sich daher auf die Übertragung von 40 Sätzen in den Dialekt seiner Erhebungsorte. Sein Motto war, „lieber *Weniges* aus möglichst *allen*, als *Vieles* aus einer *ungenügenden* Zahl an Ortschaften einzusammeln" (Wenker 1881: VIII; im Orig. Hervorhebung durch Sperrung). Die Datenerhebung erfolgte indirekt durch den Versand von Fragebogen an alle Schulorte des damaligen Deutschen Reichs. Die Lehrer wurden gebeten, gemeinsam mit den Schülern, „die obenstehenden Fragen gewissenhaft zu beantworten sowie eine Uebersetzung der einliegenden hochdeutschen Sätzchen in die ortsübliche Mundart umstehend einzutragen" (vgl. Fleischer 2017: 188). Insgesamt lagen Wenker nach Abschluss seiner Erhebung 48.505 (deutsche und fremdsprachliche) Fragebogen vor, von denen allerdings knapp 2.500 als „unbrauchbar" bezeichnet wurden (vgl. Wenker 2013: 2). Für den Sprachatlas wurden 46.011 Fragebogen aus 42.496 Erhebungsorten verarbeitet (vgl. Wenker 2013: 9). Bei den Fragebogen handelt es sich um eine geschlossene Sammlung von Texten einer Kategorie. Die Texte sind als Transkriptionen gesprochener Sprache durch die Informanten, Lehrer und Schüler, zu bewerten. Derzeit liegen die Transkriptionen lediglich in handschriftlicher Form vor. Obwohl in dem Projekt *Digitaler Wenker-Atlas* (DiWA) zwischen 2001 und 2009 alle Wenkerkarten und die Original-Erhebungsbogen digitalisiert und im Internet zur freien Verfügung bereitgestellt wurden,[4] sind bisher nur die kartierten Teile des Korpus nach Kartenthemen durchsuchbar. Die Fragebogen sind lediglich als Bilddateien zugänglich, eine Volltextdigitalisierung steht dagegen noch

4 Seit 2010 sind alle Karten des *Digitalen Wenker-Atlas* über die Forschungsplattform https://www.regionalsprache.de (letzter Zugriff: 31. 7. 2017) abrufbar.

aus. Die Karten sind darüber hinaus durch ein morphologisches (vgl. Rabanus 2005), ein phonetisch-phonologisches (vgl. Lameli 2009) und ein historisches Register erschlossen (siehe auch Abschnitt 5).

Die erste Auswertung des Korpus haben Wenker und seine Kollegen selbst vorgenommen, indem sie auf seiner Basis einen Sprachatlas mit 1.668 Karten(blättern) zu 576 Kartenthemen erstellt haben. Darüber hinaus hat Wenker „eine zum Buch gebundene 59-seitige Einleitung zum Sprachatlas des Deutschen Reichs, 431 erläuternde Texthefte im Umfang von ca. 2600 Seiten zu 356 wortbezogenen Karten des Atlasses sowie ein separates Textheft zu den Sprachverhältnissen im Nordosten des Untersuchungsgebietes" (Wenker 2013: XVII) verfasst. Weitere Auswertungen im Sinne von raumbezogenen Korpusanalysen, die sich unter anderem auf die über den Sprachatlas erschlossenen Korpusteile stützen, sind Ferdinand Wredes Dialekteinteilung (vgl. Wrede 1937), Peter Wiesingers *Phonetisch-phonologische Untersuchungen zur Vokalentwicklung in den deutschen Dialekten* (Wiesinger 1970) sowie sein Beitrag „Die Einteilung der deutschen Dialekte" (Wiesinger 1983). Echte raumstatistische Auswertungen des Wenkermaterials hat – für das Gebiet der Bundesrepublik Deutschland – erstmals Alfred Lameli (v. a. Lameli 2013) vorgelegt. Als Grundlage für diese Analysen hat er für ein Set von 66 lautlichen und morphologischen Variationsphänomenen die Varianten für die 439 Landkreise Deutschlands in einer Datenbank erfasst. Auf dieser Datenbasis wurden nicht nur raumstatistische linguistische Analysen durchgeführt, sondern auch die Sprachdaten in ihrem Verhältnis zu außersprachlichen Daten betrachtet (vgl. dazu Heblich; Lameli & Riener 2015; Falck, Lameli & Ruhose 2016). Jürg Fleischer führt außerdem derzeit auf Basis der Wenkerbogen ein Projekt *Morphosyntaktische Auswertung von Wenkersätzen* durch, für das erste Ergebnisse bereits publiziert wurden (vgl. etwa Fleischer 2012).

2.2 Mittelrheinischer Sprachatlas (MRhSA)

Von 1978 bis 1988 wurden Sprachdaten für den von Günter Bellmann initiierten Mittelrheinischen Sprachatlas (Bellmann, Herrgen & Schmidt 1994–2002) erhoben.[5] Das Untersuchungsgebiet umfasst das linksrheinische Rheinland-Pfalz und das Saarland, dialektgeographisch also Teile des Rhein- und des Moselfränkischen. Bei dem Atlas handelt es sich um den ersten bidimensionalen Dialektatlas für das Deutsche. Es wurden Sprachaufnahmen mit älteren Landwirten (um 75 Jahre) – als Datenserie 1 – und mit Handwerkern der mittleren

---

5 Vgl. zum MRhSA zusammenfassend auch Girnth (2015).

Generation (um 35 Jahre) – als Datenserie 2 – erhoben, die für die Ausübung ihres Berufes täglich von ihrem Wohnort in eine andere Ortschaft im Nahverkehrsbereich pendeln (vgl. Bellmann 1994: 40). Bellmann formuliert die Ziele des MRhSA folgendermaßen:

> Er will einmal für ein Teilgebiet der deutschen Sprachfläche – als Laut- und Formenatlas – in der üblichen Weise die Verteilung der Dialektmerkmale auf der lokal-sprachlandschaftlichen (horizontalen) Ebene dokumentieren. Indem er dies tut – darin besteht sein zweites Ziel –, richtet er seinen Blick auch auf die zwischen Dialekt und Standardsprechsprache sich erstreckende vertikale Dimension, deren Reflexe von den Sprachatlanten des Deutschen sonst, wenn überhaupt, dann eher als Störfaktoren wahrgenommen worden sind. (Bellmann 1994: 1)

Die Sprachdaten wurden in direkter Befragung mit jeweils mehreren Informanten pro Ort (549 Orte für Datenserie 1 und 292 Orte für Datenserie 2) erhoben und dabei sowohl auf Tonbändern aufgezeichnet als auch unmittelbar als IPA-Transkriptionen in Formulare eingetragen. Das Korpus besteht insgesamt aus 1.100 Abfragewörtern für die Datenserie 1 und 440 Abfragewörtern für die Datenserie 2.

Im Unterschied zu vielen anderen Dialektatlanten ist das für den Atlas erhobene Korpus im Rahmen des REDE-Projekts durch orthographische Transkriptionen der Erhebungskontexte und eine Ton-Text-Synchronisierung[6] (Alignierung) aufbereitet worden und somit zugänglich und weiter analysierbar (siehe auch Abschnitt 5). Insgesamt sind 714 transkribierte und nach Freitext durchsuchbare Sprachaufnahmen verfügbar.

2.3 Bayerische Dialektdatenbank (BayDat)

Den Bayerischen Sprachatlas (BSA) charakterisiert Horst Haider Munske (2015) unter anderem als „(a) ein dialektologisches Großprojekt mit sechs Standorten in Bayern [...], (b) eine Sammelpublikation von über 50 großformatigen Sprachatlas-Bänden [...] und (c) eine einheitliche, systematische, wissenschaftlich fundierte Exploration der Dialekte im Freistaat Bayern, deren Ergebnisse in enger Transkription im Internet (unter BayDat) zugänglich sind“ (Munske 2015: 1). Im vorliegenden Beitrag geht es also vor allem um den zuletzt genannten Aspekt. Die sechs Teilprojekte des Bayerischen Sprachatlas sind:

---

6 Die Synchronisierung von Ton und Transkription stellt für Korpora gesprochener Sprache gegenüber der reinen Transkription für viele Fragestellungen eine erhebliche Steigerung des möglichen Erkenntnisgewinns dar (vgl. auch Hunston 2008: 159–160).

- Sprachatlas von Bayerisch-Schwaben (SBS; bearbeitet in Augsburg),
- Sprachatlas von Unterfranken (SUF; bearbeitet in Würzburg),
- Sprachatlas von Mittelfranken (SMF; bearbeitet in Erlangen),
- Sprachatlas von Niederbayern (SNiB; bearbeitet in Passau),
- Sprachatlas von Nordostbayern (SNOB; bearbeitet in Bayreuth) und
- Sprachatlas von Oberbayern (SOB; bearbeitet in Passau).

Zwischen 1981 und 1998 wurden in allen Teilprojekten in insgesamt 1.613 Erhebungsorten ältere, ortsfeste und dialektkompetente Informanten in einer direkten Befragung aufgenommen. Die Fragebücher umfassten jeweils bis zu 2.800 Fragen. Die Antworten der Informanten wurden direkt in Teuthonista (vgl. Teuchert 1924/1925) phonetisch transkribiert, teilweise erfolgten zusätzlich Tonbandaufnahmen. Die Transkriptionen wurden nachträglich in den Teilprojekten digital erfasst.

Am Würzburger Standort des Bayerischen Sprachatlas erfolgte dann die Integration der digital erfassten Transkriptionen aller Standorte in eine gemeinsame Datenbank, die *Bayerische Dialektdatenbank* (BayDat; vgl. http://www.baydat.uni-wuerzburg.de:8080/cocoon/baydat/ [letzter Zugriff: 31. 7. 2017]). Diese Datenbank ermöglicht es den Benutzern, über Suchanfragen „eigene Auswertungen vorzunehmen, z. B. Belegorte, sachliche oder sprachliche Themen auszuwählen und eigene Sprachkarten zu entwerfen" (Munske 2015: 11). Die Ausgabe der Suchanfragen erfolgt jeweils in Form der Teuthonista-Transkription im HTML- oder im PDF-Format. Auch hier ist ein deutlicher Mehrwert gegenüber den rein in Atlasbänden publizierten Dialektatlanten erreicht worden, indem der große Teil der Daten, die nicht auf Karten abgebildet werden konnten, zugänglich und auswertbar ist. Die Daten der BayDat sind auch über die Internetplattform *Regionalsprache.de* in Form von über 4.000 Karten abrufbar. Es handelt sich dabei entsprechend der Datengrundlage um Vollformenkarten, auf denen die Transkription am Ort wiedergegeben ist.

Die Daten, die im Rahmen des Bayerischen Sprachatlas erhoben wurden, sind – wie in allen Sprachatlasprojekten – zunächst einmal für die Kartenerstellung ausgewertet worden. Darüber hinaus wurden mit den Daten des Sprachatlas von Bayerisch-Schwaben im Rahmen des DFG-Projekts *Neue Dialektometrie mit Methoden der stochastischen Bildanalyse* raumstatistische Analysen durchgeführt (vgl. zusammenfassend Pröll et al. 2015).

3 Variationslinguistische Tonkorpora

3.1 Variationslinguistische Tonkorpora am IDS

Das größte Korpus am IDS und zugleich das größte variationslinguistische Tonkorpus für den deutschen Sprachraum ist das *Zwirner-Korpus* (ZW). Eberhard Zwirner verfolgte das Ziel, von allen deutschen Dialekten Tonaufnahmen zu machen und diese wissenschaftlich auszuwerten (vgl. Bethge 1976: 25). Neben der vollständigen Erfassung der deutschen Dialekte war Zwirner ebenso an der sprachlichen Dynamik und weiteren Formen des regionalen Sprechens interessiert, was sich in der Anlage seiner Erhebung äußert (s. Fußnote 8, vgl. auch Schmidt & Herrgen 2011: 119–121). Für das Vorhaben wurden zwischen 1951 und 1972 an ca. 1.000 Orten in den alten Bundesländern, im Elsass, den Niederlanden, Liechtenstein und Vorarlberg Sprachaufnahmen durchgeführt. Das Gebiet wurde in Planquadrate (10 Längen- und 15 Breitenminuten, vgl. Zwirner & Bethge 1958: 15) aufgeteilt und in jedem der Quadrate wurde ein möglichst vom Verkehr abgeschiedenes Dorf bzw. in manchen Fällen auch eine kleinere Stadt ausgewählt (vgl. Zwirner & Bethge 1958: 16).[7] In jedem Untersuchungsort wurden sechs Sprecher aufgezeichnet: drei ortsfeste einheimische Sprecher und drei Heimatvertriebene. Bei den autochthonen Sprechern wurden jeweils ein Sprecher im Alter von über 60 Jahren aus der „ländlichen Bevölkerung" (Zwirner & Bethge 1958: 17), ein Sprecher der mittleren Generation und ein Sprecher im Alter um die 20 Jahre ausgewählt.[8] Bei den Heimatvertriebenen wurden die Sprecher der am häufigsten vertretenen landsmannschaftlichen Gruppe berücksichtigt (vgl. dazu Zwirner & Bethge 1958: 16–17). Durchgeführt wurden die Aufnahmen durch Germanisten der Landesuniversitäten, Wörterbuchleiter und Volkskundler (vgl. zur genauen Übersicht Zwirner & Bethge 1958: 13–14), die von technischen Mitarbeitern unterstützt wurden. Aufgenommen wurden Erzählungen der Gewährspersonen, teilweise auch Dialoge, die Abfrage der Wochentage und Zahlen (Eins bis Zehn) im Dialekt und teilweise auch die Abfrage der Wenkersätze im Dialekt (vgl. Zwirner & Bethge 1958: 18–19). Bei den Aufnahmen zielte Zwirner auf eine möglichst natürliche Gesprächssituation

---

7 Um Herford verdichten sich die Aufnahmeorte, da auf Wunsch des Westfälischen Heimatbunds im damaligen Landkreis Herford alle Orte aufgenommen wurden (insges. 180). Für eine weitere Verdichtung von Tonaufnahmen in Anlehnung an die Zwirner-Erhebung sorgte Arno Ruoff mit Aufnahmen im Südwesten des deutschen Sprachraums (vgl. Bethge 1976: 27, sowie Tab. 6.1).

8 Durch diese Auswahl versprach sich Zwirner, „sprachliche Veränderungen, insbesondere den Übergang zum Hochdeutschen […] zu erfassen" (Zwirner & Bethge 1958: 17).

und die Verwendung der Sprache, „in der der Sprecher sich in seinem Haus oder an seinem Arbeitsplatz zu unterhalten pflegt“ (Zwirner & Bethge 1958: 19). Mitunter wurden die Sprecher, soweit es den Aufnahmeleitern möglich war, auch im jeweiligen Dialekt befragt (vgl. Zwirner & Bethge 1958: 15). Nach den Aufzeichnungen wurde ein Aufnahmeprotokoll mit zahlreichen Metadaten (u. a. auch eine Einordnung der Sprachschicht und der Sprechsituation) angefertigt (vgl. Zwirner & Bethge 1958: 21–24). Auf diese Weise entstanden insgesamt 5.796 Sprachaufnahmen im Umfang von 1.077 Stunden. Durch die Gegensätzlichkeit der „objektiven sprachgebrauchssteuernden Faktoren und [des] geforderte[n] Sprachverhalten[s]“ (Schmidt & Herrgen 2011: 121) lässt sich davon ausgehen, dass die Sprecher bei den Aufnahmen variiert haben bzw. variieren mussten. Dies hat wiederum zur Folge, dass das Zwirner-Korpus „für die regionalsprachliche Sprachdynamikforschung [eine] sehr ergiebigen Datenserie“ (Schmidt & Herrgen 2011: 121) darstellt.[9]

Auch für das Gebiet der ehemaligen DDR liegt ein umfassendes variationslinguistisches Tonkorpus vor: „Deutsche Mundarten DDR“ (DR, oder kurz: DDR-Korpus). Der Begründer, Hans-Joachim Schädlich, verfolgte in Zusammenarbeit mit dem Institut für deutsche Sprache und Literatur der Akademie der Wissenschaft der DDR das Ziel der systematischen Erfassung der Dialekte sowie der „landschaftlich bestimmten Formen der Umgangssprache“ (Schädlich & Eras 1964: 376) in der damaligen DDR. Für dieses Vorhaben wurde das Gebiet in Planquadrate (10 Längen- und 15 Breitenminuten, vgl. Schädlich & Eras 1964: 377) aufgeteilt und in jedem dieser Planquadrate ein Ortspunkt ausgewählt (insgesamt 437 Ortspunkte, vgl. Schädlich & Eras 1965: 24). Bei der Auswahl, die den Aufnahmeleitern (Mitarbeiter der Wörterbücher, vgl. Schädlich & Eras 1964: 376–377) oblag, wurden Orte berücksichtigt, „bei denen mit günstigen sprachlichen Verhältnissen gerechnet werden konnte“ (Schädlich & Eras 1964: 377). Aufgezeichnet wurden pro Ort drei ortsfeste Sprecher der bodenständigen Bevölkerung aus drei Generationen (vgl. Schädlich & Eras 1964: 377). Dabei wurden verschiedene Situationen erhoben. Zum einen mussten die Informanten Vergleichstexte in den Dialekt übertragen. Diese bestanden aus den jeweiligen Wortlisten und festen Texten der Mundartwörterbücher sowie einer Variation der Wenkersätze. Die Texte wurden den Sprechern zur Vorbereitung vorab zugeschickt, teilweise wurde auch eine schriftliche Erhebung durchgeführt (vgl. dazu Schädlich & Eras 1964: 377). Zusätzlich wurden freie – aber gelenkte – Gespräche und Erzählungen in einer intendierten freien und ungezwungenen Gesprächsatmosphäre (vgl. Schädlich & Eras 1964: 376–377) aufgenommen. Nach den Aufnahmen wurde durch den Leiter ein ausführliches Auf-

9 Die Archivierung und Zugänglichkeit der Korpora wird in Abschnitt 5 beschrieben.

nahmeprotokoll angefertigt (vgl. Schädlich & Eras 1964: 381). Von 1960 bis 1964 wurden insgesamt 1.642 Sprachaufnahmen im Umfang von ca. 385 Stunden durchgeführt. Durch die Anlage der Erhebung kann auch der „Frage nach den sprachlichen Verhältnissen zwischen den Generationen einerseits und dem Verhältnis der Mundart und der Umgangssprache andererseits“ (Schädlich & Eras 1964: 382) nachgegangen werden.[10] Mit dem DDR-Korpus liegt somit für das Gebiet der ehemaligen DDR ein mit dem Zwirner-Korpus vergleichbares Korpus vor.

Mit den *Tonaufnahmen der Vertriebenenmundarten* (TAVM-Korpus, auch *Deutsche Mundarten: ehemalige Ostgebiete* [OS] genannt) existiert ein weiteres umfangreiches Variationskorpus für den deutschen Sprachraum. Zwar lagen mit der Erhebung Zwirners und der Ostdeutschen Dialektgeographie (siehe Abschnitt 3.2) schon erste Aufnahmen und Vorarbeiten zu den Dialekten der Vertriebenen vor, doch sah Bellmann (1964: 64) damals die „eingehende Erfassung der sterbenden ostdeutschen Dialekte [...] in zwölfter Stunde“ geboten. Konkretes Ziel des Vorhabens war es, die bedrohten standardfernsten Sprechlagen (d. h. die Dialekte bzw. die „Umgangssprache der untersten Stufe“ [Bellmann 1964: 66]) der Sprecher aus den ehemaligen deutschen Ostgebieten systematisch zu erfassen. Somit lässt sich von einer „archaisierenden Gesamtzielsetzung“ (Bellmann 1970: 10) sprechen. Durchgeführt wurde das Vorhaben in Zusammenarbeit verschiedener Institutionen (u. a. Deutscher Sprachatlas, Deutsches Spracharchiv, vgl. Bellmann 1964: 64). Dazu wurden zwischen 1962 und 1965 an den damaligen Wohnorten (ca. 900 Orte) insgesamt 981 Sprachaufnahmen im Gesamtumfang von 462 Stunden durchgeführt. Es wurden Sprecher ausgewählt, die über 60 Jahre alt waren, aus bäuerlichen oder ähnlichen Schichten kamen und vor ihrer Umsiedlung ortsfest waren (vgl. dazu und zur allgemein schwierigen Sprecherauswahl und -suche Bellmann 1964: 67–70). Mit ihnen wurden folgende Aufnahmen durchgeführt: Abfrage der leicht modifizierten Wenkersätze (vgl. Bellmann 1970: 24–27), der Wochentage und Zahlen (Eins–Fünfzehn) im Dialekt und ein freier Text, der in den meisten Fällen einem Monolog mit thematischer Vorgabe entspricht (vgl. Bellmann 1964: 73–76). Schmidt & Herrgen (2011: 123) weisen dem TAVM-Korpus einerseits einen „extrem hohen dokumentarischen Wert [...] für die Lauterscheinungen der untergegangenen Varietäten“ zu. Andererseits eignet sich das Korpus auch für verschiedene sprachdynamische Analysen (vgl. Schmidt & Herrgen 2011: 123).[11]

10 Im Arbeitsgebiet des Thüringischen Wörterbuchs wurde zusätzlich explizit die „geläufige Form der Umgangssprache aufgezeichnet“ (Schädlich & Eras 1964: 377).

11 Bellmann (1964: 70–72) selbst beschreibt sprachdynamische Prozesse zur Zeit der Aufnahmen. So wurden bspw. stigmatisierte Sprechweisen vermieden oder so hat bspw. für die Oberschlesier das Sorbische die Funktion des Dialekts übernommen.

Mit den drei variationslinguistischen Tonkorpora (Zwirner, DDR, TAVM) ist für den Zeitraum zwischen 1955 und 1965 für den deutschen Sprachraum „in ausgezeichneter Qualität die Dialektperformanz bzw. die stark regional geprägte Alltagssprache dreier Altersgruppen mit bis zu sechs Sprechern aus verschiedenen sozialen Schichten dokumentiert“ (Schmidt & Herrgen 2011: 123). Dies gilt weitgehend auch für die Dialektkompetenz.

Zusätzlich liegt für den Zeitraum mit dem *Pfeffer-Korpus* auch ein Tonkorpus der Umgangssprache vor. Alan Pfeffer vom Institut für Grunddeutsch der Universität Pittsburgh setzte sich zum Ziel, die gesprochene Umgangssprache der deutschsprachigen Länder zu Anfang der 1960er Jahre zu dokumentieren (vgl. Pfeffer & Lohnes 1984: 9; oft ist auch die Rede von gesprochener Alltagssprache, städtischer oder gehobener Umgangssprache, vgl. Pfeffer 1975: 28; zum Begriff Grunddeutsch, vgl. Pfeffer 1975: 5–10). In Zusammenarbeit mit verschiedenen Institutionen in den entsprechenden Ländern (BRD, DDR, Österreich, Schweiz, vgl. Pfeffer & Lohnes 1984: 14) entstanden 1961 insgesamt 401 Sprachaufnahmen in einem Umfang von ca. 79 Stunden. Die Aufzeichnungen wurden in 56 Städten durchgeführt, wobei sich die Anzahl der Aufnahmen pro Stadt nach der Bevölkerungsstatistik richtete (vgl. Pfeffer & Lohnes 1984: 13–14). Erhoben wurden 185 Sprecherinnen und 218 Sprecher (insgesamt 403 Sprecher), deren Auswahl nach Bildung, Beruf, Alter und Größe des Wohnorts variiert wurde (vgl. Pfeffer & Lohnes 1984: 17; Pfeffer 1975: 28–29). Von den Sprechern wurden hauptsächlich ungezwungene dialogische Gespräche, teilweise auch Monologe mit Einleitung durch den Aufnahmeleiter und Gruppengespräche aufgezeichnet (vgl. Pfeffer 1975: 11, 29). Dafür wurden 25 Themen systematisch über die Aufnahmen verteilt (vgl. dazu Pfeffer & Lohnes 1984: 17–18). Aufgrund des Erkenntnisinteresses „gesprochene Alltagssprache“ wurden zu stark dialektale Sprachproben aus dem Korpus ausgeschlossen (vgl. Pfeffer 1975: 11). Insgesamt zeichnet sich das Korpus gemäß dem Ziel bzw. der Anlage der Erhebung durch eine relativ große sprachliche Variation aus (vgl. Spiekermann 2008: 107), wird aber meist als standardnäheres Korpus bewertet (vgl. Lenz 2007: 179).

Die beschriebenen variationslinguistischen Tonkorpora sind von großem Nutzen für die Validierung von Ergebnissen und Daten (vgl. Schmidt & Herrgen 2011: 115, 123), wurden bisher aber selten umfassend systematisch ausgewertet. Zu nennen sind bei den Auswertungen der Korpora die Phonai-Monographien zum Zwirner-Korpus (vgl. http://pub.ids-mannheim.de/abgeschlossen/phonai/ [letzter Zugriff: 31. 7. 2017]), mit denen für 30 Ortspunkte phonetische Transkriptionen und phonologische Analysen vorliegen. Für den ersten Band des *Schlesischen Sprachatlasses* (Schmitt 1965–1967) hat Bellmann die entsprechenden Tonaufnahmen der Vertriebenenmundarten ausgewertet. Der *Datenbank regionaler Umgangssprachen des Deutschen* (DRUGS) liegen u. a. Tonauf-

Tab. 6.1: Weitere variationslinguistische Tonkorpora in der DGD2.

Korpus	**Böblingen-Korpus (BB)**	**Südwestdeutschland-Vorarlberg-Korpus (SV)***	**Schwarzwald-Korpus (SW)***	**König-Korpus (KN)**
Sprechlage	v. a. intendierter Dialekt	intendierter Dialekt	v. a. intendierter Dialekt	Vorlese-aussprache
Sprecher	Sprecher verschiedenen Alters	Gewährspersonen des *Vorarlberger Sprach-atlas* (VALTS; vgl. Gabriel 1985)	Sprecher ab 5 Jahren	43 Akademiker (17–27 Jahre)
Region	Kreis Böblingen	Südwestdeutschland, Liechtenstein, Vorarlberg	drei Weiler im Schwarzwald	43 Ortspunkte in den alten Bundesländern[12]
Aufnahme-datum	1965–1967	1966–1970	1964/1974	1975–1976
Aufnahme-situation	Erzählungen, Gespräche & Vorlese-aussprache	Erzählungen, Wochentage, Zahlen	Erzählungen, Wochentage, Zahlen & Vorlese-aussprache	Vorlese-aussprache (Ausschnitt des Grundgesetzes)
Verfügbar-keit	73 Aufnahmen (insges. 42,5h) in DGD2	242 Aufnahmen (insges. 72h) in DGD2	126 Aufnahmen (insges. 37,5h) in DGD2	43 Aufnahmen (insges. 42,5h) und Transkripte in DGD2

* angelehnt an die Zwirner-Erhebung

nahmen des Pfeffer-Korpus zugrunde (vgl. Lauf 1994: 6). Als raumbezogene Korpusanalyse ist Spiekermann (2008) zu nennen, der in seiner Untersuchung zur Sprache in Baden-Württemberg die entsprechenden Tonaufnahmen des Pfeffer-Korpus ausgewertet hat (vgl. Spiekermann 2008: 105–107). Das DDR-Korpus wurde u. a. für den *Sprachatlas für Rügen und die vorpommersche Küste* (Herrmann-Winter 2013) analysiert. Zudem liegen zahlreiche einzelthematische Untersuchungen vor, die sich auf die genannten Korpora stützen. Als Beispiele können Rowley (1997), Fleischer (2002), Freywald (2010), Lenz (2013), Werth (2014) oder Kasper & Werth (2015) genannt werden, die vorwiegend syntaktische Auswertungen mit den Daten des Zwirner- und/oder des Pfeffer-Korpus vornehmen.

12 Der *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland* (König 1989) basiert auf Aufnahmen mit Sprechern aus 44 Orten, von denen 43 Aufnahmen im Korpus enthalten sind.

3.2 Weitere variationslinguistische Korpora

Für den deutschen Sprachraum liegen weitere variationslinguistische Tonkorpora vor, die hier in Auswahl gesondert behandelt werden, da sie nicht Teil des IDS/der DGD2 sind. Hierzu zählt die „Ostdeutsche Dialektgeographie“. Walther Mitzka und Ludwig Erich Schmitt verfolgten das Ziel der dialektgeographischen Erfassung der ehemaligen deutschen Ostgebiete (siehe Abschnitt 3.1). Im Rahmen des Vorhabens wurden von 1954 bis 1958 mit Sprechern aus diesen Gebieten an ihren damaligen Wohnorten Tonaufnahmen durchgeführt. Aufnahmeleiter waren dabei Mitarbeiter des Deutschen Sprachatlas, der Wörterbuchprojekte, der Landsmannschaften, Lehrer und interessierte Laien (vgl. Göschel 1977: 15). Aufgezeichnet wurden die Dialektabfrage der Wenkersätze, der Wochentage und Zahlen sowie Erzählungen und teilweise Dialoge (vgl. Göschel 1977: 14). Die Angaben zur Anzahl der Aufnahmen differieren. Es wird angenommen, dass ca. 1.050–1.100 Aufnahmen angefertigt wurden (vgl. Göschel 1977: 15). Davon sind 729 Wenkersatz-Aufnahmen im REDE-System zugänglich (siehe Abschnitt 5). Göschel (1977: 15) merkt zudem an, dass durch die Entstehung des Korpus die „Qualität der Aufnahmen und [die] räumliche Verteilung sehr uneinheitlich“ sind.

In Vorbereitung des Hessischen Dialektzensus wurde 1982 die Mundart-Aktion „Ich sag’s hessisch“ von den hessischen Sparkassen, der Landesbausparkasse und der Hessen-Nassauischen Versicherung in Zusammenarbeit mit dem Hessen-Nassauischen Wörterbuch (unter Leitung von Heinrich Dingeldein) durchgeführt. Daraus entstanden ist das Korpus *Tonaufnahmen der hessischen Mundarten* (TAHM). Die Sprecher sollten eine Kassette in Selbstaufnahme besprechen. Auf der A-Seite sollten die Wenkersätze in den Dialekt übersetzt werden, die B-Seite stand für freie Texte, Lieder und Inszenierungen zur Verfügung. Insgesamt liegen 1.200 Tonaufnahmen hauptsächlich von Schülerinnen und Schülern, aber auch von Sprechern der mittleren und älteren Generation vor. Davon sind 508 Wenkersatz-Aufnahmen im REDE-System zugänglich (siehe Abschnitt 5). Durch die Anlage der Erhebung sind zum Teil nur eingeschränkt Metadaten vorhanden. Zudem ist bei der Aufnahme der Dialekt zwar intendiert, doch ist vorwiegend von Regiolektaufnahmen auszugehen (persönliche Auskunft von H. Dingeldein; vgl. zu diesem Abschnitt auch https://regionalsprache.de/tonkorpora.aspx#TAHM [letzter Zugriff: 31. 7. 2017]). In Anlehnung an das TAHM-Korpus wurde von 2003 bis 2004 unter Leitung von Stefan Arend das Projekt *Tonarchiv osthessischer Mundarten* (ToM) durchgeführt, das zum gleichnamigen Korpus führte. Auch hier führten hauptsächlich Sprecherinnen und Sprecher der älteren und mittleren Generation Selbstaufnahmen (Wenkersätze, Erzählungen und Gedichte) in 23 Gemeinden und Ortsteilen des Landkreises Fulda durch. Insgesamt liegen 400 Tonaufnahmen

vor, von denen 179 Wenkersatz-Aufzeichnungen im REDE-System zugänglich sind. Alle Aufnahmen sind auch über das Medienzentrum Fulda (http://tom.medienzentrum-fulda.de [letzter Zugriff: 31. 7. 2017]) abrufbar (vgl. hierzu https://regionalsprache.de/tonkorpora.aspx#TOM [letzter Zugriff: 31. 7. 2017] sowie Arend 2002/2003).[13]

Im Kontext variationslinguistischer Tonkorpora ist auch das aktuelle Forschungsprojekt *Sprachalltag II: Sprachatlas – Digitalisierung – Nachhaltigkeit* an der Tübinger Arbeitsstelle Sprache in Südwestdeutschland – Arno-Ruoff-Archiv zu erwähnen. Eines der Ziele des Projekts besteht in der Digitalisierung der ca. 2.000 Tonaufnahmen des Arno-Ruoff-Archivs und der dazu erstellten Transkriptionen im Umfang von etwa 20.000 Seiten (vgl. dazu http://www.wiso.uni-tuebingen.de/faecher/empirische-kulturwissenschaft/ta-sprache/forschung/sprachalltag-ii-sprachatlas-digitalisierung-nachhaltigkeit.html [letzter Zugriff: 31. 7. 2017]).

4 Moderne variationslinguistische Korpora

Die Korpora, die unter diesem Punkt zusammengefasst werden, zeichnen sich aus durch (a) moderne Erhebungsmethoden und (b) eine umfassende Perspektivierung sprachlicher Variation, die auch eine Gesamtbetrachtung der modernen Regionalsprachen ermöglicht (vgl. Schmidt & Herrgen 2011). Hinsichtlich der Dimensionen sprachlicher Variation werden sowohl die diatopische (horizontale) als auch die vertikale und die diachrone Dimension betrachtet.

Ein modernes variationslinguistisches Korpus (*Deutsch heute*) ist im Rahmen des IDS-Projekts *Variation des gesprochenen Deutsch* entstanden.[14] Ziel des Projekts – in Anschluss an das Pfeffer- und das König-Korpus (siehe Abschnitt 3.1) – war die „Beschreibung der sprachlichen Variation in [formellen] Situationen“ (Kleiner 2015: 491) und letztlich die Beantwortung der Frage, wie die Standardsprache verwendet wird. Dazu wurden von 2006 bis 2009 an 192 Ortspunkten im gesamten deutschsprachigen Raum Aufnahmen durchgeführt (Deutschland 146, Österreich 25, Schweiz 13, Italien 3, Luxemburg 2, Belgien 2, Liechtenstein 1; vgl. Kleiner 2015: 492 und http://prowiki.ids-mannheim.de/bin/view/AADG/OrtsListe [letzter Zugriff: 31. 7. 2017]). Aufgezeich-

13 Für die beiden Korpora TAHM und ToM gilt, dass die Aufnahmen durch die Entstehung der Korpora, deren Vorteile auf praktischer Seite liegen, in Bezug auf akustische Qualität und sprachliche Realisierung recht heterogen sind (vgl. https://regionalsprache.de/tonkorpora.aspx [letzter Zugriff: 31. 7. 2017]).

14 Vgl. zum *Deutsch heute*-Korpus zusammenfassend auch Kleiner (2015).

net wurden 345 weibliche und 326 männliche Sprecher mit höherer Schulbildung im Alter von 16 bis 20 Jahren (d. h. Oberstufenschüler) sowie als Vergleichsgruppe 81 weibliche und 77 männliche Sprecher im Alter von 50 bis 60 Jahren. Folgende Situationen wurden erhoben: Vorleseaussprache („Nordwind und Sonne“ [Fabel], populärwissenschaftlicher Text, konstruierte Texte), Übersetzung (Englisch > Deutsch), Bildbeschreibung, Wortliste (Standard), sprachbiographisches Interview und Wegbeschreibung (Map Task). Insgesamt liegen 829 Aufnahmen im Umfang von ca. 1.244 Stunden vor. Diese wurden orthographisch transkribiert und (teilweise automatisch) aligniert. Ausschnitte der Aufnahmen wurden des Weiteren phonetisch transkribiert. Im Rahmen des *Forschungs- und Lehrkorpus gesprochenes Deutsch* (FOLK) als Teil der DGD2 sind einige Interviews und Map Tasks des Korpus bereits zugänglich. Mittelfristig ist geplant, das gesamte Korpus in die DGD2 aufzunehmen. Ergebnisse des Projekts werden u. a. online als *Atlas zur Aussprache des deutschen Gebrauchsstandards* (AADG) (Kleiner 2011 ff.) publiziert.

Eine weitere umfassende Analyse der sprachlichen Variation unternimmt das SiN-Projekt (*Sprachvariation in Norddeutschland*),[15] das von 2007 bis 2013 von der Deutschen Forschungsgemeinschaft gefördert wurde. Ziel des Projekts war die „Dokumentation und Interpretation der aktuellen Sprachsituation im Norden Deutschlands“ mit einem „integrative[n], multiperspektivische[n] Ansatz“ (Elmentaler et al. 2015: 397). Dazu wurde von 2007 bis 2010 das SiN-Korpus erstellt. An 36 Ortspunkten im niederdeutschen Sprachraum (jeweils zwei Orte in den traditionell abgegrenzten niederdeutschen Dialektgebieten) wurde die Erhebung durchgeführt (vgl. Elmentaler et al. 2015: 399–400). Dabei wurden ländlich geprägte Gemeinden und Ortsteile mit 2.000–8.000 Einwohnern gewählt. Aufgenommen wurden Sprecherinnen im Alter von 40 bis 60 Jahren.[16] Pro Erhebungsort wurden vier Sprecherinnen aufgenommen (insgesamt 144 Sprecherinnen), von denen jeweils zwei dialektkompetent sein sollten und zwei kein Niederdeutsch beherrschen sollten. Mit ihnen wurden folgende Aufnahmen durchgeführt: Dialektabfrage der Wenkersätze, (dialektale) Erzählung, informelles Tischgespräch, formelles, leitfadengestütztes Interview, Vorleseaussprache („Nordwind und Sonne“, Zeitungsartikel). Außerdem wurden mit allen Informantinnen ein Salienz-, Situativitäts-, Normativitäts- und Arealitätstest durchgeführt. Pro Sprecherin dauerten die Aufnahmen ca. 3 Stunden, sodass ein Korpus im Umfang von ca. 432 Stunden vorliegt. Die festen Texte sowie Ausschnitte (2.500 Wörter) von Tischgesprächen und

15 Vgl. zu SiN zusammenfassend auch Elmentaler et al. (2015).

16 Zur komplementären Auswahl in Bezug zum REDE-Projekt, vgl. Elmentaler et al. (2015: 399) sowie Ganswindt, Kehrein & Lameli (2015: 431).

Interviews wurden normorthographisch bzw. nach Niederdeutschkonventionen transkribiert und aligniert. Ausschnittweise wurden auch phonetische Transkriptionen angefertigt. Einige der Tischgespräche sind in der DGD2 (FOLK) zugänglich. Das Gesamtkorpus wird am Hamburger Zentrum für Sprachkorpora (HZSK) dauerhaft gesichert und „soll mittelfristig der wissenschaftlichen Forschung zur Verfügung gestellt werden“ (Elmentaler et al. 2015: 411). Die Ergebnisse liegen vor und werden sukzessive publiziert (vgl. Elmentaler et al. 2015: 411–424).

Das Forschungsprojekt *Regionalsprache.de* (REDE) wird seit 2008 am Forschungszentrum Deutscher Sprachatlas der Universität Marburg durchgeführt.[17] Gefördert wird das Projekt, dessen Laufzeit 19 Jahre beträgt, durch die Akademie der Wissenschaften und der Literatur Mainz. Das übergeordnete Ziel des Projekts ist die umfassende und systematische Erforschung der modernen Regionalsprachen des Deutschen. Für dieses übergeordnete Ziel sind zwei konkretere Teilziele formuliert: (1) Aufbau eines forschungszentrierten Informationssystems zu den modernen Regionalsprachen des Deutschen (siehe dazu Abschnitt 5) sowie (2) Ersterhebung und Analyse der variationslinguistischen Struktur und Dynamik der modernen Regionalsprachen des Deutschen. Für das zweite Teilziel wurde unter anderem ein umfangreiches modernes variationslinguistisches Korpus (REDE-Korpus) erstellt. Von 2008 bis 2015 wurden Sprachaufnahmen an 150 Orten (Städte und Gemeinden) im gesamten Gebiet der Bundesrepublik Deutschland durchgeführt (vgl. https://regionalsprache.de/empirie-stand-der-bearbeitung.aspx [letzter Zugriff: 31. 7. 2017]). Aufgenommen wurden jeweils mindestens vier ortsfeste Sprecher aus drei Generationen. Die Sprecher der älteren Generation waren über 65 Jahre alt und sind (meist) einer handwerklichen oder landwirtschaftlichen Tätigkeit nachgegangen. In der mittleren Generation wurden 45- bis 55-jährige Polizisten aufgezeichnet und schließlich in der jüngsten Generation 18- bis 23-jährige Abiturienten aus dem jeweiligen Ort (insgesamt über 700 Sprecher, da an manchen Orten aus forschungspraktischen Gründen mehr als die vorgesehenen Repräsentanten einer Generation aufgezeichnet wurden). Die Erhebungssituationen des REDE-Projektes versuchen, verschiedene Grade der Formalität und somit verschiedene Grade der sprecherseitigen Orientierung an der Standardsprache zu evozieren. Folgende Aufnahmen wurden durchgeführt: Abfrage der Wenkersätze im Dialekt und im individuell besten Hochdeutsch, Vorleseaussprache („Nordwind und Sonne“), formelles sprachbiographisches Interview sowie informelles Gespräch mit einem vertrauten Gesprächspartner ohne Anwesenheit Dritter

17 Vgl. zum REDE-Projekt zusammenfassend auch Ganswindt, Kehrein & Lameli (2015).

(sog. Freundesgespräch). Die Aufnahmen dauerten jeweils ca. 2,5 Stunden, sodass bei einer Gesamtzahl von 4.026 Aufnahmen ein Korpus im Umfang von ca. 1.633 Stunden vorliegt. Die Aufbereitung der Daten erfolgt sukzessive. Die standardisierten Texte werden vollständig und die freien Situationen auszugsweise normorthographisch und feinphonetisch transkribiert sowie mit dem Tonsignal synchronisiert. Die aufbereiteten Daten und Forschungsergebnisse werden sukzessive im REDE-System (siehe Abschnitt 5) veröffentlicht. Momentan sind 2.014 Aufnahmen und Transkripte zugänglich.

5 Zugänge zu Dialekt- und Variationskorpora

Die vier in Abschnitt 3.1 beschriebenen variationslinguistischen Tonkorpora des IDS (Zwirner, DDR, TAVM, Pfeffer) sind ins *Deutsche Spracharchiv* (DSAv) und später ins *Archiv für gesprochenes Deutsch* (AGD) übergegangen und liegen vollständig digitalisiert in der *Datenbank für gesprochenes Deutsch* (DGD2) vor. Diese ist seit 2012 das Nachfolgesystem der DGD1. Aufgrund der zentralen Bedeutung der Korpora und der Datenbank für die Variationslinguistik sollen der Zugang und die Nutzung kurz umrissen werden (vgl. zum gesamten Abschnitt Schmidt, Dickgießer & Gasch 2013).[18] Die DGD2 ist ein Korpusmanagementsystem, in dem Teilbestände des ADG (Sprachaufnahmen, Transkripte, Metadaten) zur Verfügung gestellt werden (zu den anderen Korpora der DGD2 siehe Abschnitt 3.1, Boas & Fingerhuth in diesem Band sowie Schmidt in diesem Band und Internetseite der DGD). Die Sprachaufnahmen liegen digitalisiert und teilweise tontechnisch bearbeitet im WAVE-Format vor. Die Transkripte sind größtenteils ebenfalls digitalisiert bzw. digital erstellt worden. Durch die unterschiedliche Bearbeitung der Transkripte liegt hier eine große Heterogenität vor, an deren Auflösung stetig gearbeitet wird. Zusätzlich steht eine umfangreiche Korpusdokumentation zur Verfügung, die zahlreiche Metadaten zu Aufnahmen, Sprechern usw. enthält. Nach einmaliger Online-Registrierung kann die DGD2 genutzt werden. Je nach Erkenntnisinteresse bestehen unterschiedliche Zugriffs- und Bearbeitungsformen. Die DGD2 bietet zwei grundlegende Funktionen an: Browsen und Suchen. „Der Menüpunkt ‚Korpora' der DGD2 bietet die Möglichkeit, Metadaten, Transkripte und Zusatzmaterial anzusehen sowie Audioaufnahmen ausschnittsweise anzuhören." (Schmidt, Dickgießer & Gasch 2013: 13) Dies wird als Browsen verstanden und ermöglicht

18 Zusätzlich können die Sprachaufnahmen der Korpora auch im persönlichen Service über die DGD2 bestellt werden bzw. bietet die DGD2 auch einen Download-Service an.

dem Nutzer eine erste Orientierung in den Korpora. Im Vergleich zur gezielten Suche entspricht diese Funktion eher einem „Stöbern" und wird durch verschiedene Formen der Verknüpfungen (Links) zusätzlich verbessert. Beim Browsen lassen sich jeweils 15-sekündige Ausschnitte anhören und, sofern das Transkript vorliegt, auch die Verschriftlichungen mitlesen (vgl. Schmidt, Dickgießer & Gasch 2013: 14–16). Bei der gezielten Suche lassen sich zwei Funktionen unterscheiden. Bei der Volltextsuche wird innerhalb der Transkripte nach dem jeweiligen Begriff gesucht, wobei über Wildcards usw. Spezifizierungen möglich sind. Die struktursensitive Suche wiederum greift nicht nur auf die Transkripte, sondern auch auf die Annotationen zurück und ermöglicht zudem über Metadatenfilter eine Spezifizierung der Fundstellen (vgl. Schmidt, Dickgießer & Gasch 2013: 17–19). Mit der DGD2 – seit April 2017 steht die Version 2.8 mit einigen Verbesserungen und Neuerungen zur Verfügung – liegt somit eine stabile, stets aktualisierte Online-Datenbank vor, die sich nicht nur durch die umfangreichen Bestände, sondern auch durch Zugänglichkeit und Funktionalität auszeichnet und somit für die variationslinguistische Forschung eine wichtige Datenquelle darstellt.

Wie oben erwähnt besteht eines der Teilziele im Projekt REDE im Aufbau einer interaktiven Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Diese Plattform ist in den letzten Jahren auf Basis des Datenbestandes aus dem *Digitalen Wenker-Atlas* aufgebaut worden und sie wird stetig erweitert.[19] Das Herzstück des REDE-Systems bildet das sprachgeographische Informationssystem REDE SprachGIS, über das kartenbasiert auf den gesamten Datenbestand zugegriffen werden kann. Um dies zu ermöglichen, sind alle Daten des Systems orts- oder raumbezogen gespeichert und mit einer eindeutigen geographischen Identifikationsnummer (GID) versehen, die auf geographische Koordinaten verweist. Teil des Datenbestandes bilden auch die im vorliegenden Beitrag beschriebenen oder genannten Sprachatlas- und Tonkorpora.

Die kartierten Phänomene aus den im System enthaltenen Sprachatlaskorpora sind über die Kartensuche aufrufbar. In dieser Kartensuche kann neben der Freitexteingabe auch eine Filterung über ein morphologisches, ein phonetisch-phonologisches und ein historisches Register vorgenommen werden. Karten können dann einzeln oder übereinander in das REDE SprachGIS geladen und so betrachtet und direkt verglichen werden.

19 Die Forschungsplattform ist frei unter der URL www.regionalsprache.de (letzter Zugriff: 31. 7. 2017) zugänglich.

Alle anderen Daten, also auch die orthographisch transkribierten Teile der Tonkorpora,[20] können über das Recherche-Werkzeug des REDE SprachGIS durchsucht werden. Die Suche kann dabei über alle Korpora erfolgen oder sie kann auf einzelne oder mehrere ausgewählte Korpora eingeschränkt werden. Zusätzlich besteht die Möglichkeit, Suchergebnisse über einen Kartenelementefilter auf eine ausgewählte Region (z. B. ein Dialektverband, ein Bundesland oder auch ein vorher frei gezeichnetes Polygon) zu begrenzen. Diese Recherche- und Filterfunktionen stehen nicht nur für die Tonkorpora, sondern für alle im System enthaltenen Daten zur Verfügung, z. B. auch für das Wenkerbogen-Korpus (vgl. Kehrein i. Dr.). Für dieses Korpus kann bisher allerdings nur angezeigt werden, an welchen Orten ein Bogen verfügbar ist. Eine Suche in den Bogen ist (noch) nicht möglich, sie liegen lediglich als Bilddateien vor. Die Suchergebnisse für die Tonkorpora lassen sich im REDE SprachGIS einzeln oder als Zusammenstellung („Playlist“) anhören und auf diese Weise vergleichend analysieren.

Die beiden beschriebenen Suchwerkzeuge (Kartensuche und Recherche) und ihre je spezifischen Filterfunktionen werden derzeit in einem übergreifenden Suchwerkzeug zusammengeführt. Das bedeutet, dass nach erfolgter Umstellung auch bei der Suche in den entsprechend erschlossenen Tonkorpora nach den Kategorien der vorhandenen Register gefiltert werden kann.[21]

6 Ausblick

Die vorhandenen Korpora zur Sprachvariation bzw. zu sprachlichen Varietäten im Deutschen bilden bereits eine solide und umfangreiche Grundlage für wissenschaftliche Studien. Dennoch könnte und sollte diese Basis in den

20 Außer den im vorliegenden Beitrag beschriebenen Korpora sind im REDE-System noch Aufnahmen aus weiteren Korpora enthalten. Da es sich dabei um einen Bestand handelt, der bereits für den *Digitalen Wenker-Atlas* zusammengestellt wurde, sind im REDE-System meist Wenkersatzaufnahmen aus den Korpora integriert (vgl. für eine Übersicht Kehrein i. Dr. sowie https://regionalsprache.de/tonkorpora.aspx [letzter Zugriff: 31. 7. 2017]).

21 Aktuell zu nennen ist auch [moca] (*multimodal oral corpora administration*). Hierbei handelt es sich um ein Online-System der Universität Freiburg (Romanisches Seminar), das Tonaufnahmen mit alignierten Transkripten und Metadaten speichert und zur Verfügung stellt. Es werden umfangreiche und detaillierte Suchmöglichkeiten sowie die Funktion der individuellen Bearbeitung der Korpora angeboten. Das System ist ohne Einschränkungen online zugänglich und soll in der dritten Version ab November 2017 zur Verfügung stehen (vgl. http://www.hpsl.uni-freiburg.de/ [letzter Zugriff: 31. 7. 2017]).

kommenden Jahren sowohl in quantitativer als auch in qualitativer Hinsicht erweitert werden. Quantitativ sind folgende Maßnahmen denkbar:

– Erhöhung der Ortsnetzdichte der modernen variationslinguistischen Korpora durch (regionale) Neuerhebungen mit vergleichbaren Methoden,
– Ausbau vorhandener Datensammlungen, z. B. von Dialektatlanten, zu analysierbaren Korpora – ähnlich wie beim Mittelrheinischen Sprachatlas oder der Bayerischen Dialektdatenbank – durch Transkriptionen, durch Ton-Text-Synchronisierung, durch Erschließung über Register und evtl. durch Datenbankerfassung der Transkriptionen,
– Volltextdigitalisierung von handschriftlichen (Korpus-)Transkriptionen, die bisher nicht oder nur in Bildform digital vorliegen (z. B. das Wenkerbogen-Korpus im REDE-System oder die in Tübingen bearbeiteten Ruoff-Transkriptionen),
– Integration solcher neuen Korpora in vorhandene Datenbanksysteme oder zumindest Vernetzung mit diesen.

In qualitativer Hinsicht könnten vorhandene und neue Dialekt- und Variationskorpora nicht nur durch orthographische und – wie teilweise bereits durchgeführt – phonetische Transkriptionen erschlossen werden. Es könnten vielmehr kontrollierte Annotationen für alle linguistischen Systemebenen vorgenommen werden, was die (technisch bereits möglichen) Analyseoptionen um ein Vielfaches erhöhen würde. Wichmann (2008) bringt dieses Desiderat folgendermaßen auf den Punkt:

> One hopes that future spoken corpora will provide linguistically sophisticated syntactic, pragmatic and discourse annotation together with an equally sophisticated prosodic annotation that can then be complemented by automatic analysis of global trends, such as pitch, pause, loudness and voice quality. At present, the technology outstrips the linguistics. (Wichmann 2008: 205)

Literatur

Arend, Stefan (2002/2003): Reden, wie der Schnabel gewachsen ist. Tonarchiv osthessischer Mundarten (ToM) will Dialekte der Region umfassend und flächendeckend dokumentieren. *Jahrbuch Landkreis Fulda* 30, 124–138.

Auer, Peter (1990): *Phonologie der Alltagssprache. Eine Untersuchung zur Standard/Dialekt-Variation am Beispiel der Konstanzer Stadtsprache*. Berlin, New York: de Gruyter.

Bellmann, Günter (1964): Wege und Möglichkeiten der Schallaufnahme ostdeutscher Mundarten heute. Zur Tonbandaufnahme der Vertriebenenmundarten. *Zeitschrift für Mundartforschung* 31, 62–79.

Bellmann, Günter (1970): Einleitung. In Günther Bellman & Joachim Göschel (Hrsg.), *Tonbandaufnahmen ostdeutscher Mundarten 1962–1965. Gesamtkatalog mit 10 Karten*, 7–29. Marburg: Elwert.
Bellmann, Günter (1994): *Einführung in den Mittelrheinischen Sprachatlas*. Tübingen: Niemeyer.
Bellmann, Günter, Joachim Herrgen & Jürgen Erich Schmidt (Hrsg.) (1994–2002): *Mittelrheinischer Sprachatlas*. 5 Bde. Tübingen: Niemeyer.
Bethge, Wolfgang (1976): Vom Werden und Wirken des Deutschen Spracharchivs. *Zeitschrift für Dialektologie und Linguistik* 43 (1), 22–43.
Chambers, J. K. & Peter Trudgill (1998): *Dialectology. Second edition*. Cambridge: Cambridge University Press.
Christen, Helen, Ingrid Hove & Marina Petkova (2015): Gesprochene Standardsprache im Deutschschweizer Alltag. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 379–396. Berlin, Boston: de Gruyter Mouton.
Christen, Helen, Manuela Guntern, Ingrid Hove & Marina Petkova (2010): *Hochdeutsch in aller Munde. Eine empirische Untersuchung zur gesprochenen Standardsprache in der Deutschschweiz* (Zeitschrift für Dialektologie und Linguistik. Beiheft 140). Stuttgart: Steiner.
Coseriu, Eugenio (1992): *Einführung in die Allgemeine Sprachwissenschaft*. 2. Auflage. Tübingen: Francke.
Digitaler Wenker-Atlas (DiWA) (2001–2009). Hrsg. von Jürgen Erich Schmidt & Joachim Herrgen. Bearbeitet von Alfred Lameli/Alexandra N. Lenz/Jost Nickel & Roland Kehrein/Karl-Heinz Müller/Stefan Rabanus. Marburg: Forschungszentrum Deutscher Sprachatlas. http://www.diwa.info (letzter Zugriff: 31. 7. 2017).
Elmentaler Michael, Joachim Gessinger, Jens Lanwer, Peter Rosenberg, Ingrid Schröder & Jan Wirrer (2015): Sprachvariation in Norddeutschland (SiN). In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 397–424. Berlin, Boston: de Gruyter Mouton.
Falck, Oliver, Alfred Lameli & Jens Ruhose (2016): Cultural biases in migration: Estimating non-monetary migration costs. *Papers in Regional Science*, doi: 10.1111/pirs.12243.
Fleischer, Jürg (2002): *Die Syntax von Pronominaladverbien in den Dialekten des Deutschen. Eine Untersuchung zu Preposition Stranding und verwandten Phänomenen*. Stuttgart: Steiner.
Fleischer, Jürg (2012): Pronominalsyntax im nordwestlichen Niederdeutsch: eine Auswertung des Wenker-Materials (mit Einbezug der friesischen und dänischen Formulare). *Niederdeutsches Jahrbuch* 135, 59–80.
Fleischer, Jürg (2017): *Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen. Dokumentation, Entdeckungen und Neubewertungen*. Hildesheim u. a.: Olms.
Fleischer, Jürg & Thomas Gadmer (Hrsg.) (2002): *Schweizer Aufnahmen. Tondokumente aus dem Phonogrammarchiv der Österreichischen Akademie der Wissenschaften. Gesamtausgabe der Historischen Bestände 1899–1950*. Wien: Österreichische Akademie der Wissenschaften.
Freywald, Ulrike (2010): Obwohl vielleicht war es ganz anders. Vorüberlegungen zum Alter der Verbzweitstellung nach subordinierenden Konjunktionen. In Arne Ziegler (Hrsg.), *Historische Textgrammatik und Historische Syntax des Deutschen*, 55–84. Berlin, Boston: de Gruyter.

Friginal, Eric & Jack A. Hardy (2014): *Corpus-based sociolinguistics. A guide for students*. New York, London: Routledge.

Gabriel, Eugen (Hrsg.) (1985): *Vorarlberger Sprachatlas mit Einschluss des Fürstentums Liechtenstein, Westtirols und des Allgäus (VALTS). Einführung in den Vorarlberger Sprachatlas*. Bregenz: Vorarlberger Landesregierung.

Ganswindt, Brigitte, Roland Kehrein & Alfred Lameli (2015): Regionalsprache.de (REDE). In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 425–453. Berlin, Boston: de Gruyter Mouton.

Girnth, Heiko (2015): Der Mittelrheinische Sprachatlas (MRhSA). In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 29–51. Berlin, Boston: de Gruyter Mouton.

Göschel, Joachim (Hrsg.) (1977): *Die Schallaufnahme deutscher Dialekte im Forschungsinstitut für deutsche Sprache*. Marburg.

Heblich, Stephan, Alfred Lameli & Gerhard Riener (2015): The impact of regional accents on economic behavior: A lab experiment on linguistic performance, cognitive ratings and economic decisions. *PLoS ONE* 10/2, doi: 10.1371/journal.pone.0113475.

Herrmann-Winter, Renate (2013): *Sprachatlas für Rügen und die vorpommersche Küste*. Kartographie Martin Hansen. Rostock: Hinstorff.

Hunston, Susan (2008): Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics. An international handbook*. Vol. 1 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), 154–168. Berlin, Boston: de Gruyter Mouton.

Jaberg, Karl & Jakob Jud (1928): *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz*. Halle (Saale): Niemeyer.

Kasper, Simon & Alexander Werth (2015): Fundierung linguistischer Basiskategorien (LingBas). In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 349–377. Berlin, Boston: de Gruyter Mouton.

Kehrein, Roland (2006): Regional accent in the German language area – How dialectally do German police answer emergency calls? In Frans Hinskens (Hrsg.), *Language variation. European perspectives*, 83–96. Amsterdam, Philadelphia: Benjamins.

Kehrein, Roland (2012): *Regionalsprachliche Spektren im Raum. Zur linguistischen Struktur der Vertikale*. Stuttgart: Steiner.

Kehrein, Roland (i. Dr.): Das Wenker-Material im REDE SprachGIS. In Jürg Fleischer et al. (Hrsg.), *Minderheitensprachen und Sprachminderheiten: Deutsch und seine Kontaktsprachen in der Dokumentation der Wenker-Materialien*. Hildesheim u. a.: Olms.

Kleiner, Stefan (2011 ff.): *Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG)*. Unter Mitarbeit von Ralf Knöbl. http://prowiki.ids-mannheim.de/bin/view/AADG/ (letzter Zugriff: 31. 7. 2017).

Kleiner, Stefan (2015): „Deutsch heute“ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 489–518. Berlin, Boston: de Gruyter Mouton.

König, Werner (1989): *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland*. 2 Bde. Ismaning: Hueber.

König, Werner (Hrsg.) (1996–2009): *Sprachatlas vom Bayerisch-Schwaben* (Bayerischer Sprachatlas: Regionalteil 1). 14 Bde. Heidelberg: C. Winter.

Kunst, Jan Pieter & Sjef Barbiers (2010): Generating maps on the internet. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (Hrsg.), *Language and space. An international handbook of linguistic variation. Volume 2: Language mapping* (Handbücher zur Sprach- und Kommunikationswissenschaft 30.2), 401–415. Berlin, Boston: de Gruyter Mouton.

Lameli, Alfred (2004): *Standard und Substandard. Regionalismen im diachronen Längsschnitt*. Stuttgart: Steiner.

Lameli, Alfred (2009): Kommentar zum phonologischen Register der standardsprachlichen Vergleichslaute. http://www.regionalsprache.de/phonologisches-register.aspx (letzter Zugriff: 31. 7. 2017).

Lameli, Alfred (2013): *Strukturen im Sprachraum. Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Berlin, Boston: de Gruyter.

Lauf, Raphaela (1994): *Datenbank regionaler Umgangssprachen des Deutschen. DRUGS. Abschlussbericht*. Manuskript. Universität Marburg.

Lehmann, Christian (2007): Daten – Korpora – Dokumentation. In Werner Kallmeyer & Gisela Zifonun (Hrsg.), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*, 9–27. Berlin, Boston: Walter de Gruyter.

Lenz, Alexandra N. (2003): *Struktur und Dynamik des Substandards. Eine Studie zum Westmitteldeutschen (Wittlich/Eifel)*. Stuttgart: Steiner.

Lenz, Alexandra N. (2007): Zur variationslinguistischen Analyse regionalsprachlicher Korpora. In Werner Kallmeyer & Gisela Zifonun (Hrsg.), *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*, 187–202. Berlin, Boston: de Gruyter.

Lenz, Alexandra N. (2013): *Vom kriegen und bekommen. Kognitiv-semantische, variationslinguistische und sprachgeschichtliche Perspektiven*. Berlin, Boston: de Gruyter.

Munske, Horst Haider (2015): Der Bayerische Sprachatlas (BSA). In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 1–27. Berlin, Boston: de Gruyter Mouton.

Pfeffer, J. Alan (1975): *Grunddeutsch: Erarbeitung und Wertung dreier deutscher Korpora; ein Bericht aus dem Institute for Basic German, Pittsburgh*. Tübingen: Narr.

Pfeffer, J. Alan & Walter F. W. Lohnes (Hrsg.) (1984): *Grunddeutsch. Texte zur gesprochenen deutschen Gegenwartssprache*. Tübingen: Niemeyer.

Pröll, Simon, Simon Pickl, Aaron Spettl, Volker Schmidt, Evgeny Spodarev, Stephan Elspaß & Werner König (2015): Neue Dialektometrie mit Methoden der stochastischen Bildanalyse. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (Hrsg.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 173–194. Berlin, Boston: de Gruyter Mouton.

Rabanus, Stefan (2005): Kommentar zum morphologischen Register. http://www.regionalsprache.de/morphologisches-register.aspx (31. 7. 2017).

Regionalsprache.de (REDE) (2008 ff.). Hrsg. von Jürgen Erich Schmidt, Joachim Herrgen & Roland Kehrein. Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Bearbeitet von Dennis Bock, Brigitte Ganswindt, Heiko Girnth, Simon Kasper, Roland Kehrein, Alfred Lameli, Slawomir Messner, Christoph Purschke, Anna Wolańska. Marburg: Forschungszentrum Deutscher Sprachatlas.

Rocholl, Josephine (2015): *Ostmitteldeutsch* – eine *moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum*. Hildesheim u. a.: Olms.

Rowley, Anthony Robert (1997): *Morphologische Systeme der nordostbayerischen Mundarten in ihrer sprachgeographischen Verflechtung*. Stuttgart: Steiner.

Schädlich, Hans-Joachim & Heinrich Eras (1964): Deutsche Dialektologie und Tonaufnahmetechnik. *Spektrum. Mitteilungsblatt für die Mitarbeiter der Deutschen Akademie der Wissenschaften zu Berlin* 10, 375–382.

Schädlich, Hans-Joachim & Heinrich Eras (1965): Bericht über die Tonbandaufnahmen der deutschen Mundarten in der Deutschen Demokratischen Republik. In *Berichte über dialektologische Forschungen in der Deutschen Demokratischen Republik*, 24–27. Berlin.

Schmidt, Jürgen Erich & Joachim Herrgen (2011): *Sprachdynamik. Eine Einführung in die moderne Regionalsprachenforschung*. Berlin: Erich Schmidt.

Schmidt, Thomas, Sylvia Dickgießer & Joachim Gasch (2013): *Die Datenbank für Gesprochenes Deutsch – DGD2*. Mannheim: Institut für Deutsche Sprache.

Schmitt, Ludwig Erich (Hrsg.) (1965–1967): *Schlesischer Sprachatlas*. Band 1: Laut- und Formenatlas. Von Günter Bellmann. Unter Mitarbeit von Wolfgang Putschke und Werner Veith. Band 2: Wortatlas. Von Günter Bellmann. Marburg: Elwert.

Spiekermann, Helmut (2008): *Sprache in Baden-Württemberg: Merkmale des regionalen Standards*. Tübingen: Niemeyer.

Teuchert, Hermann (1924/1925): Lautschrift des Teuthonista. In *Teuthonista* 1, 5.

Viëtor, Wilhelm (1888): Beiträge zur Statistik der Aussprache des Schriftdeutschen. *Phonetische Studien. Zeitschrift für wissenschaftliche und praktische Phonetik* I, 95–114; 209–226.

Vorberger, Lars (2017): *Regionalsprache in Hessen. Eine Untersuchung zu Sprachvariation und Sprachwandel im mittleren und südlichen Hessen*. Dissertationsschrift Universität Marburg.

Wecker-Kleiner, Bernadette (2009): *Sprechen nach der Schrift. Die Vorleseaussprache von DialektsprecherInnen in Bayerisch-Schwaben im Spannungsfeld zwischen Dialekt und Orthoepie*. Berlin: dissertation.de.

Wenker, Georg (1881): *Sprach-Atlas von Nord- und Mitteldeutschland. Auf Grund von systematisch mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30 000 Orten. Bearbeitet, entworfen und gezeichnet von Dr. G. Wenker. Text. Einleitung*. Straßburg, London: Trübner.

Wenker, Georg (1886): [Über das Sprachatlasunternehmen]. In *Verhandlungen der achtundreißigsten Versammlung deutscher Philologen und Schulmänner in Gießen vom 30. September bis 3. Oktober 1885*, 187–194. Leipzig: Teubner.

Wenker, Georg (1889–1923): *Sprachatlas des Deutschen Reichs*. Handgezeichnet von Emil Maurmann, Georg Wenker & Ferdinand Wrede. Forschungszentrum Deutscher Sprachatlas. Marburg. [Publiziert als *Digitaler Wenker-Atlas* (DiWA)].

Wenker, Georg (2013): *Schriften zum Sprachatlas des Deutschen Reichs. Gesamtausgabe. Band 1: Handschriften: Allgemeine Texte, Kartenkommentare 1889–1897*. Hrsg. und bearb. von Alfred Lameli. Unter Mitarbeit von Johanna Heil und Constanze Wellendorf. Hildesheim u. a.: Olms.

Werth, Alexander (2014): Die Funktionen des Artikels bei Personennamen im norddeutschen Sprachraum. In Friedhelm Debus, Rita Heuser & Damaris Nübling (Hrsg.), *Linguistik der Familiennamen*, 139–174. Hildesheim u. a.: Olms.

Wichmann, Anne (2008): Speech corpora and spoken corpora. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics. An international handbook.* Vol. 1 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), 187–207. Berlin, Boston: de Gruyter Mouton.

Wiesinger, Peter (1970): *Phonetisch-phonologische Untersuchungen zur Vokalentwicklung in den deutschen Dialekten*. 2 Bände (Studia Linguistica Germanica 2/1+2). Berlin, New York: de Gruyter.

Wiesinger, Peter (1983): Die Einteilung der deutschen Dialekte. In Werner Besch et al. (Hrsg.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. 2. Teilband (Handbücher zur Sprach- und Kommunikationswissenschaft 1.2), 807–900. Berlin, New York: de Gruyter.

Wrede, Ferdinand (1937): Ferdinand Wredes Einteilungskarte der deutschen Mundarten. In Walther Mitzka & Bernhard Martin (Hrsg.), *Deutscher Sprachatlas. 9. Lieferung. Karte 56*. Marburg: Elwert.

Zwirner, Eberhard & Wolfgang Bethge (1958): *Erläuterung zu den Texten* (Lautbibliothek der deutschen Mundarten 1). Göttingen: Vandenhoeck & Ruprecht.

Hans C. Boas und Matthias Fingerhuth

7 Deutsche Sprachinselkorpora im 21. Jahrhundert

Abstract: Dieser Beitrag umreißt die Dokumentation von Sprachinseln des Deutschen in ihrer historischen Entwicklung, ihren gegenwärtigen Stand sowie die Herausforderungen, die sich bei der Erhaltung und Verfügbarmachung dieser Daten für ein (Fach-)Publikum stellen. Nachdem die Sprachinselforschung anfänglich auf geschriebenen Notizen basierte, benutzte sie über weite Teile des 20. Jahrhunderts analoge Medien. Die analog gespeicherten Daten sind gefährdet, sowohl durch den Verfall der Medien selbst als auch durch den der Abspielgeräte. Darüber hinaus schränken die analogen Medien Zugang zu und Vervielfältigung der Daten stark ein. Durch Digitalisierung kann nicht nur die langfristige Archivierung gewährleistet, sondern auch ein vereinfachter globaler Zugang gewährt werden. Obwohl somit technische Lösungen für die dringendsten Probleme der analogen Aufnahmen existieren, gibt es weitere Hürden für die Erstellung zugänglicher Archive. Die Sprachinselforschung hat sich über lange Zeit weitgehend dezentral entwickelt, und einzelne Forschungsprojekte haben jeweils eigene Methodologien und Forschungsinfrastrukturen verwendet. In manchen Fällen sind erhobene Daten verschollen, auf der anderen Seite sind nicht sämtliche erhaltenen Daten dokumentiert. Darüber hinaus geben rechtliche Vorgaben gegenwärtig einen engen Rahmen für die Archivierung und Nutzung von Sprecherdaten ohne die explizite Einwilligung der Sprecher vor, die jedoch nicht bei allen existierenden Aufnahmen vorliegt. Trotz dieser Hürden gibt es Bemühungen, die Archivierung und Veröffentlichung historischer Sprachinseldaten zu verbessern, und die Erhebung neuer Daten zu koordinieren. Nennenswerte Beispiele dafür sind die *Datenbank für Gesprochenes Deutsch* am Institut für Deutsche Sprache in Mannheim sowie das *Sprachinselarchiv* an der University of Texas at Austin.

Anmerkung: Für hilfreiche Kommentare bedanken wir uns bei Csaba Földes, Nicole Eller-Wildfeuer, Claudia Riehl, Peter Wagener und den Herausgebern dieses Bandes.

Hans C. Boas, Department of Germanic Studies, The University of Texas at Austin, 1 University Station, C3300, 2505 University Blvd., Austin, Texas 78712, U.S.A., E-Mail: hcb@mail.utexas.edu
Matthias Fingerhuth, Institut für Germanistik, Universität Wien, Universitätsring 1, A-1010 Wien, Österreich, E-Mail: matthias.fingerhuth@univie.ac.at

https://doi.org/10.1515/9783110538649-008

Keywords: Dialektologie, Spracharchiv, Sprachdokumentation, Sprachinsel, Sprachkontakt

1 Einleitung

In diesem Beitrag wird über Korpora berichtet, mit denen deutsche Sprachinseldialekte untersucht werden können. Der Schwerpunkt liegt auf digitalen Sprachinselkorpora, die seit Ende des 20. Jahrhunderts in der sprachwissenschaftlichen Forschung immer mehr an Bedeutung gewonnen haben, gerade auch im Rahmen einer intensiveren Beschäftigung mit empirischen Methoden und Einsichten der Korpuslinguistik (Lüdeling & Kytö 2008/2009). Dieser Beitrag ist wie folgt gegliedert: Abschnitt 2 gibt einen historischen Überblick über die Dokumentation und Archivierung von Sprachinseldaten.[1] Abschnitt 3 diskutiert zunächst technische und methodologische Aspekte der Erstellung von digitalen Sprachinselkorpora. Anschließend wird ein Überblick über eine Reihe unterschiedlicher digitaler Sprachinselkorpora gegeben. Abschnitt 4 diskutiert methodologische und technische Aspekte einer vergleichenden Sprachinselforschung mit digitalen Sprachinselkorpora und zeigt, wie gewisse sprachwissenschaftliche Probleme mit neuen technischen Methoden gelöst werden können. Abschnitt 5 fasst den Beitrag zusammen.

2 Historischer Überblick

2.1 Datenerhebung

Das im Laufe des 19. Jahrhunderts aufkommende Interesse an der Erforschung deutscher Dialekte führte zu unterschiedlichen Arten der Datenerhebung. Die womöglich bekannteste Art der damaligen Datenerhebung lässt sich auf Georg Wenker zurückführen, der 1878 den weltweit ersten Sprachatlas, den *Sprach-Atlas der Rheinprovinz nördlich der Mosel sowie des Kreises Siegen* veröffentlichte. Dieser Sprachatlas bildete die Grundlage für den späteren *Sprachatlas des Deutschen Reichs*, bestehend aus 1.668 von Hand gezeichneten Karten.

Die Datenerhebung selbst bestand aus der Versendung von Fragebögen mit 40 volkstümlichen Sätzen (die sogenannten Wenkersätze), die Wenker mit der

1 Zur Begriffsbestimmung des Begriffs „Sprachinsel“ siehe Wildfeuer (2017a).

Hilfe von Lehrern in die jeweiligen Ortsdialekte übersetzen ließ (indirekte Erhebung). Die zurückgeschickten Fragebögen dienten als Datenbasis für den Sprachatlas, in dem auf Teilkarten die erhobenen Daten eingetragen wurden. Später wurden Nacherhebungen für die Gebiete außerhalb des Deutschen Reiches vorgenommen, so dass für Europa eine nahezu vollständige Erhebung, einschließlich einiger deutscher „Sprachinseln", vorliegt. Insgesamt konnten 51.480 Bögen aus 49.363 deutschsprachigen Orten gesammelt werden, zusätzlich gingen 2.050 fremdsprachige Bögen ein. Alle Originalbögen sind im Forschungsinstitut „Deutscher Sprachatlas" in Marburg archiviert (Herrgen & Lenz 2003). Die von Wenker verwendeten 40 Sätze bildeten im 20. Jahrhundert auch die Basis für viele Sprachatlanten der sogenannten Marburger Schule, so dass es schrittweise gelang, die deutschen Dialekte mittels der Marburger Erhebungsmethode in einer gewissen Vollständigkeit zu erfassen (Knoop, Putschke & Wiegand 1982; Goebl & Schiltz 2006). Der Deutsche Sprachatlas bildete die Grundlage für zahlreiche Publikationen und inspirierte im Laufe des 20. Jahrhunderts auch viele Ortsgrammatiken sowie Sprachatlanten deutscher Sprachinseln (Reed & Seifert 1954; Klein & Schmitt 1961/1965; Gilbert 1972; Brenner, Erb & Manherz 2008).

Die von O. Bremer vorgebrachte Kritik an der Methodik der Marburger Schule führte dazu, dass auch direkt erhobene, in kleinräumigen Untersuchungen zusammengestellte Daten erhoben wurden. Als solches ist die direkte Befragung von Gewährspersonen die häufigste Form der dialektologischen Datenerhebung. Sie erlaubt es, kompetente Sprecher als Experten heranzuziehen und ihnen bedachtsam ausgewählte Fragen zur Beantwortung vorzulegen (Wagener 1988: 103–104). Obwohl es an Kritik zum Vorgehen des Sprachatlasses nicht mangelt (dazu mehr im Folgenden), und zahlreiche Unsicherheiten bezüglich der so erhobenen Daten verbleiben, kann man es doch als einen methodischen Fortschritt gegenüber vorhergehender Arbeit sehen.

Als Beispiel kann man etwa Schmellers (1838) Arbeit zum Zimbrischen anführen, über deren Entstehung das Skript eines Vortrags Schmellers Hinweise birgt, jedoch keine letztliche Aufklärung verschafft. So schreibt Schmeller, dass er während einer Reise in das zimbrische Gebiet Gelegenheit hatte, „verschiedene Personen methodisch nach den Hauptrubriken ‚meines Versuches über die Mundarten Bayerns' zu vernehmen" (Schmeller 1838: 588). Die Ausführungen zur Methodologie sind in der erwähnten Schrift *Die Mundarten Bayerns* jedoch bemerkenswert kurz und beschränken sich auf den Hinweis, dass er viele der zugrunde liegenden Notizen selbst bei Wanderungen durch das Land und durch „planmäßige Vernehmung neu eingereihter Conscribierten" (Schmeller 1821: XI) erhoben hat. Aus dem Weiteren in Schmellers Vortrag Erwähnten geht jedoch auch hervor, dass Schmeller bei seiner Reise auch bemüht war, schriftliche Quellen zu kaufen oder zumindest zu kopieren.

Bei der Erhebung sprachlicher Daten wurden im vortechnischen Zeitalter traditionell nur Papier und Stift für die Erfassung sprachlicher Daten (und Metadaten) benutzt, erst später ist die Ton- und Bildaufzeichnung zu einem gängigen Werkzeug der Forschung geworden. So war z. B. in der ersten Hälfte des 20. Jahrhunderts das wissenschaftliche Interesse an der Ergründung menschlicher Sprache eine treibende Kraft für die Entwicklung der Technik für Tonaufnahmen (Schüller 2008: 4). Ungeachtet dieses Umstands und des regen Interesses für Dialektologie um die Jahrhundertwende mag es ein wenig überraschen, dass die noch junge Technik bereits Anfang des 20. Jahrhunderts zur Aufnahme von Sprachinseln eingesetzt wurde. So finden sich im Wiener Phonogrammarchiv bereits 1912 von Anton Pfalz in den Sieben Gemeinden gemachte Aufnahmen des Zimbrischen.[2]

Die Methodik der Datenerhebung wird in der frühen Sprachinselforschung teilweise nur wenig diskutiert und ist deshalb auch nur bedingt nachvollziehbar. Die Einblicke, die sich ergeben, suggerieren ein (aus heutiger Sicht) teils wenig systematisches Vorgehen. Über die von Anton Pfalz durchgeführten Aufnahmen der Zimbern in den Sieben Gemeinden erfahren wir etwa Folgendes: „Der Sprecher, [...] ein etwa fünfzigjähriger Zimber, hat die Aufnahme A und B a, b frei erfunden, B c–h nach der dort angegebenen Quelle in den Apparat gesprochen." (Lessiak & Pfalz 1918: 59) A verweist dabei auf eine Reihe von nur teilweise zusammenhängenden Sätzen, deren thematischer Schwerpunkt sich vielleicht mit ‚folkloristisch' beschreiben lässt, B auf zimbrische Sprichwörter, die meisten derer aus einer philologischen Zeitschrift entnommen sind.

Daten jüngeren Datums sind dabei nicht automatisch zeitgemäßer, wie sich, um beim Zimbrischen zu bleiben, an den von Bruno Schweizer erhobenen Daten aus den Dreizehn Gemeinden zeigt. Zur Datenerhebung schreibt er etwa:

> Ich beginne mit dem Abdruck ganz einfacher Zusammenstellungen und Redensarten, wie sie mir von meinen zimbrischen Freunden zur Einführung ihre Sprache vorgetragen oder diktiert wurden. Die Fragen und Antworten unter Nr. 9 [Gespräche] sind von meinem Herbergsvater Stefano Nordera aus einem italienisch-deutschen Sprachführer übersetzt. Leider war dieser wackere Mann von seinen täglichen Arbeiten und von der Bewirtung seiner Gäste immer so in Anspruch genommen, daß er nur selten Zeit fand, sich ein wenig für die zimbrische Sprachforschung zur Verfügung zu stellen. (Schweizer 1939: 18–19)

Für die große Mehrheit der Sprachproben ist jedoch die Quelle nicht nachvollziehbar. Zur Methode bemerkt Schweizer weiter, dass es sich „größtenteils um

2 Zu Anfang des 20. Jahrhunderts wurden an mehreren Forschungsinstitutionen Tonarchive zu Forschungszwecken gegründet: Wien (1899), Berlin (1900), St. Petersburg (1908) und Zürich (1909), vgl. Schüller (2008).

von mir selbst abgehörte und nach dem Wortlaut vom Munde des Erzählers weg in Lautschrift festgehaltene Texte" (Schweizer 1939: 9) handelt.

Die Methodenvielfalt lässt sich auch im weiteren Verlauf des 20. Jahrhunderts beobachten. Ein kurzer Einblick in die nordamerikanische Forschung kann dies, wenn nicht repräsentativ, so doch exemplarisch verdeutlichen. Die ersten sprachwissenschaftlichen Daten zum Texasdeutschen wurden in den 1930er und 1940er Jahren von Fred Eikel auf Grundlage von Übersetzungsaufgaben erhoben, bei der 191 Sätze aus dem Englischen ins Texasdeutsche übersetzt wurden (Eikel 1949). Als nächstes erhob Glenn Gilbert Anfang der 1960er Jahre Daten zum Texasdeutschen, abermals ausgehend von Übersetzungsaufgaben, jedoch auf Grundlage von Wortlisten, die er von Reed & Seifert (1954) und Atwood (1962) abgewandelt hatte (Gilbert 1963: 28–63). Dies ist insofern bemerkenswert, da auch Firchow (1991: 260) berichtet, in Minnesota auf Grundlage von Materialien zu arbeiten, die sie von Moellecken erhalten hatte, welcher sie wiederum von Reed und Seifert adaptiert hatte (cf. Moelleken 1988: 109). Hier gibt es also methodologische Überschneidungen zwischen verschiedenen Sprachinseln. Der in Fortsetzung dieser Arbeit von Gilbert (1972) erstellte *Linguistic Atlas of Texas German* basiert dagegen abermals auf der Übersetzung ganzer Sätze, die jedoch nicht mit denen Eikels übereinstimmen (Boas 2009: 9–10), und auch nicht mit der Methode von Reed und Seifert. Eine dritte Welle der Dokumentation erfolgt schließlich seit 2001 mit dem *Texas German Dialect Project* (TGDP).[3] Wie unten noch detaillierter beschrieben wird, greift das TGDP die Methode von Eikel und Gilberts Sprachatlas auf, um einen historischen Vergleich zu ermöglichen. Nimmt man jedoch weitere Projekte hinzu, so tun sich weitere Unterschiede auf. Für den *Linguistic Atlas of Kansas German*[4] wurden beispielsweise Wenkersätze sowie ein eigens konzipierter Fragebogen verwendet.

Trotz teilweiser Überschneidungen summieren sich damit die Unterschiede zwischen den verschiedenen Korpora. Eine methodologische Vielfalt ist sicher nicht grundsätzlich beklagenswert, da unterschiedliche Ansätze unterschiedlichen Erkenntniszielen entsprechen können und ein Festhalten an einer einzelnen Erhebungsweise schwer mit methodologischem Fortschritt vereinbar scheint. Es ist sicher auch so, dass sich die Verhältnisse innerhalb der verschiedenen Sprachinseln deutlich unterscheiden können, so dass Fragebögen, die auf eine Erhebung in Texas abzielen, für eine Erhebung in Wisconsin oder Sibirien ohne eine Anpassung problematisch sind. Auch das Interesse für bestimmte sprachliche Phänomene, auf deren Erkundung etwa Übersetzungs-

3 http://www.tgdp.org (letzter Zugriff: 25. 6. 2017).
4 http://www2.ku.edu/~germanic/LAKGD/Atlas_Intro.shtml (letzter Zugriff: 25. 6. 2017).

aufgaben gesondert abzielen können, kann in direkter Abhängigkeit von der Kontaktsprache stehen. Dennoch ist die bewusste Abstimmung zwischen Projekten, wie man sie in der Arbeit der letzten Jahre vermehrt beobachten kann, eine Entwicklung, die den Vergleich zwischen Erhebungen erleichtern dürfte.

2.2 Analoge Datenarchivierung und -darstellung

Ab den 1930er Jahren gab es dank technischen Fortschritts preiswertere und leichter zu benutzende Aufnahmegeräte und Mikrofone, die es den Forschern ermöglichten, mehr Tonaufnahmen durchzuführen. Neben den Tonaufnahmen fertigten Forscher auch weiterhin Fragebögen und andere schriftliche Beobachtungen an, die als Metadaten und begleitendes Forschungsmaterial mit den Tonträgern archiviert wurden. Viele Tonaufnahmen und ihre Begleitmaterialien wurden an Forschungsinstituten, Universitäten und Bibliotheken archiviert. Das wahrscheinlich bekannteste Archiv ist das Deutsche Spracharchiv, das bereits 1932 von Eberhard Zwirner in Berlin gegründet wurde, um die Analyse konstitutiver Faktoren und die Struktur gesprochener Sprache zu analysieren (Knetschke & Sperlbaum 1983). Die Tonaufnahmen wurden auf Schallplatten aufgenommen, dazu wurden Transkripte sowie Kurven und Messwerte zur phonometrischen Auswertung angefertigt (siehe Zwirner 1983 und Stift & Schmidt 2014).

Im Bereich der Sprachinselforschung gibt es das von Viktor Schirmunski an der Universität Leningrad 1927 gegründete Volksliedarchiv, dessen Daten für die Untersuchung sprachlicher Strukturen zwar nur begrenzt nützlich sind, aber dennoch ein interessantes Zeitzeugnis russlanddeutscher Sprache darstellen. Die Aufnahmen wurden von 1924–1931 aufgezeichnet und in Leningrad (heute St. Petersburg) archiviert (s. John & Swetosarowa 2005).[5] Einige Tonaufnahmen deutscher Sprachinseln finden erst Jahrzehnte später ihren Weg zu wissenschaftlichen Institutionen, wie z. B. die von Glenn Gilbert in den 1960er Jahren gemachten Tonaufnahmen, die 2006 an das Max Kade-Institut an der University of Wisconsin at Madison gegangen sind, oder die von Michael Clyne über mehrere Jahrzehnte geschaffenen Tonaufnahmen des Australiendeutschen, die 2008 an das Institut für Deutsche Sprache (IDS) in Mannheim kamen.

In den Jahren von 1991 bis 1993 hat sich das IDS um die Dokumentation der zu diesem Zeitpunkt in germanistischen Instituten, Archiven und Forschungsprojekten vorliegenden Aufnahmen des gesprochenen Deutsch bemüht,

---

5 http://www.liederlexikon.de/ueber_liederlexikon_de/projekte/bkm_projekt (letzter Zugriff: 25. 6. 2017).

indem es eine entsprechende Umfrage durchgeführt und die Ergebnisse in Folge veröffentlicht hat (Wagener & Bausch 1997). Die dort vorgenommene Erhebung weist nach Angabe der Autoren Lücken auf und war aufgrund des zeitlichen Abstandes zwischen Datenerhebung und Publikation in manchen Fällen bereits zum Veröffentlichungszeitpunkt nicht mehr aktuell, was aus heutiger Sicht umso mehr gelten muss. Dennoch stellt sie die bislang wohl vollständigste Dokumentation von Aufnahmen des Deutschen außerhalb des binnendeutschen Sprachraums dar. Einige der von Wagener & Bausch (1997) dokumentierten Aufnahmen sind mittlerweile ins Archiv für gesprochenes Deutsch aufgenommen bzw. über die Datenbank für Gesprochenes Deutsch (DGD) des IDS zugänglich gemacht worden. Obwohl analog aufgenommen, können diese Korpora damit inzwischen als digital angesehen werden, weshalb Details zum Fortschritt dieser Arbeit entsprechend weiter unten besprochen werden.

Außer solch größeren Sammlungen archivierter Tonaufnahmen ist jedoch nicht klar, wie viele Aufnahmen Forscher in ihren Privatbeständen belassen haben und diese nicht den Weg in wissenschaftliche Archive gefunden haben. Als Beispiel seien hier die Arbeiten von Fred Eikel (1949, 1954, 1966) zum New Braunfels Texas German genannt, welche auf Tonaufnahmen beruhen, die in den frühen 1940er Jahren aufgenommen wurden. Diese Tonaufnahmen wurden weder an der Johns Hopkins University, wo Eikel promoviert haben soll, noch an einer anderen Institution, wie z. B. dem New Braunfelser Sophienburg Museum archiviert. Per Zufall stießen Mitarbeiter des Texas German Dialect Projects im Herbst 2006 bei einem Interview auf eine alte Schallplatte, auf der ein von Fred Eikel 1942 aufgenommenes Interview mit einer Sprecherin des Texas Deutschen aufgezeichnet war. Diese Sprecherin erinnerte sich während des Gesprächs 2006 an die alte Schallplatte, auf der sie als neunjähriges Mädchen im Gespräch mit Eikel zu hören ist.[6] Was mit den anderen Aufnahmen von Eikel passiert ist, ist leider unbekannt, was exemplarisch zeigt, dass das Problem der fehlenden systematischen institutionellen Archivierung entstandener Sprachinseltonaufnahmen sicherlich nicht zu unterschätzen ist.

Die archivierten analogen Tonaufnahmen sowie ihre auf Papier verfassten Metadaten und andere Begleitmaterialien bilden zwar eine empirisch wertvolle Datenbasis für die Sprachinselforschung, sie sind jedoch, genau wie nur auf

6 Im Sommer 2017 fand Lars Hinrichs aus der Anglistik der University of Texas at Austin im Institut alte Tonbandbestände, die eigentlich „nur“ historische Aufnahmen des Texas Englischen sein sollten. Wie sich herausstellte, waren darunter aber auch ca. zehn Stunden historische Aufnahmen des Texasdeutschen aus den 1930er und 1940er Jahren. Leider konnte bisher nicht festgestellt werden, wer diese Aufnahmen angefertigt hatte.

Papier festgehaltene Sprachdaten wie der Deutsche Sprachatlas, aus der Benutzer-, Archivierungs- und Zugangsperspektive problematisch.[7] Erstens haben analoge Tonträger (Wachs, Platten, Magnetbänder) nur eine begrenzte Lebensdauer, je nach Qualität der Tonträger durchschnittlich 30–50 Jahre, selbst wenn diese professionell behandelt und archiviert werden.[8] Ein damit verbundenes Problem ist die Abhängigkeit von funktionierenden technischen Abspielgeräten: technische Standards ändern sich und im Laufe der technischen Entwicklung stellen Firmen die Produktion von Abspielgeräten und die Produktion von Ersatzteilen ein. Diese Entwicklung war bereits während der 1990er Jahre absehbar, wie Schüller bemerkt:

> In view of the limited life expectancy of carriers, and limited availability of replay equipment, it had become clear that the classical paradigm of museums and archives, namely to preserve the original objects or documents placed in their care, would ultimately be in vain. Long-term preservation has to concentrate on the content by extracting the signals from the original carriers, by digitising them, and by migrating these digitized contents losslessly from one IT preservation platform to the next. (Schüller 2012: 864)

Zweitens bieten analoge Sprachinseltonaufnahmen und ihre Begleitmaterialien nur eine begrenzte Möglichkeit, mit ihnen wissenschaftlich intensiv zu arbeiten. So muss man oftmals zu der Institution, an der die Sprachinseltonaufnahmen archiviert sind, reisen und dort ggf. längere Zeit verbringen. Darüber hinaus sind analoge Tonträger und deren Abspielgeräte eher mühsam zu bedienen und das wiederholte Abspielen von sensiblen Tonträgern wie Kassettenbändern kann zur starken Abnutzung und im schlimmsten Fall zu deren Unbrauchbarkeit führen. Außerdem können mehrere Wissenschaftler nicht mit denselben Tonaufnahmen gleichzeitig arbeiten, so dass gemeinschaftliche Forschungsvorhaben eher schwer zu realisieren sind.

---

7 So stellt Herrgen z. B. auch bezüglich der Zugänglichkeit und Verwendungsmöglichkeit des Deutschen Sprachatlas fest, dass eine „umfassende, auch anschauliche Version des Sprachatlasses nie publiziert werden konnte, weil die aufwendige Farbreproduktion nicht realisierbar erschien. Hier wie an mehreren anderen Stellen wirkten sich – bei einem Pionierprojekt wie dem Sprachatlas möglicherweise unvermeidbare – unglückliche methodologische Entscheidungen aus, die wiederholt, in verschiedenen Projektphasen, ohne kritischen Blick auf später sich stellende Auswertungs- und Publikationsprobleme getroffen wurden." (Herrgen 2001: 1527).

8 Ähnliche Probleme gibt es auch beim Deutschen Sprachatlas: Zum einen sind die handgezeichneten historischen Karten in ihrem Bestand gefährdet, denn die Farben (oft wurden auf ein und demselben Kartenblatt bis zu 22 Farben verwendet) beginnen zu verblassen. Zum Zweiten ist es absolut wünschenswert, die sehr anschaulichen Karten einer interessierten Öffentlichkeit zur Verfügung zu stellen. Und zum Dritten sind die wissenschaftlichen Auswertungsmöglichkeiten des Atlasses noch nicht annähernd ausgeschöpft (Herrgen & Lenz 2003).

Drittens verursachen analoge Sprachinselkorpora, wie alle Laut- und Bildarchive, erhebliche Kosten und können somit von möglichen Einsparungen oder politisch motivierten Aktionen betroffen sein. So stellt z. B. Schüller Folgendes fest:

> This often leads to dramatic situations of audiovisual collections within the realm of universities in the Western world: Their democratic and autonomous organisation is an endangering factor for the adequate or further support of audiovisual research collections, as researchers, specifically under the prevailing neo-liberal climate of our times, have an increasing tendency to optimise their short-term success at the expense of long-term strategies in the interest of the scientific community at large, including further generations to come. Additionally, such a policy implicitly takes into account that the results of present publications cannot be evaluated by researchers in the future. This would not be in line with basic scientific principles. There are several cases of audiovisual collections under the umbrella of universities whose existence is threatened. (Schüller 2008: 8)

Wenn auch die Probleme der Erstellung und langfristigen Sicherung digitaler Sprachkorpora nicht trivial sind, so lassen sich doch für diese einige Vorteile ins Feld führen. Durch den stetig sinkenden Preis von Speicherplatz lassen sich auch große Mengen von Sprachdaten verhältnismäßig günstig speichern. Ferner kann eine Vielzahl von Forschern an verschiedensten Orten parallel mit den gleichen Daten arbeiten, ohne dass vorher aufwändig Datenträger physisch vervielfältigt worden wären. Der nächste Abschnitt gibt einen Überblick über digitale Sprachinselkorpora, sowohl solche, die online über das Internet zugänglich sind, als auch solche, die „nur" vor Ort an der jeweiligen Institution genutzt werden können.

3 Digitale Sprachinselkorpora

3.1 Digitale Sprachinselkorpora am IDS

Mit dem *Archiv für gesprochenes Deutsch* verfügt das IDS Mannheim über eine bedeutende Sammlung von Sprachkorpora. In seinen Ursprüngen kann dieses zum einen in der Tradition des von Eberhard Zwirner gegründeten Deutschen Spracharchivs in die Zeit vor dem zweiten Weltkrieg zurückverfolgt werden. Dieses wurde 1971 vom IDS übernommen. Schon vor diesem Zeitpunkt hatte das IDS jedoch selbst auf die Erstellung gesprochener Sprachkorpora hingearbeitet, und auch in der Folgezeit wurden weitere Korpora erhoben (Stift & Schmidt 2014). Zum Zeitpunkt des Aufstiegs digitaler Tonaufnahmen zum Ende des 20. Jahrhunderts verfügte das IDS damit bereits über eine umfassende Sammlung von Tonaufnahmen. Im Laufe der 1990er Jahre wurde das *Deutsche*

Spracharchiv (so der damalige Name) am IDS schrittweise digitalisiert, um den Bestand der analogen Tonaufnahmen (auch von deutschen Sprachinseln) für die Zukunft zu erhalten. Ein Hauptziel war auch, durch die Verschmelzung von Ton- und Computertechnik Tonaufnahmen sehr viel schneller zugänglich zu machen, was eine einfachere Bearbeitung mit verschiedenen Instrumenten der Computertechnik ermöglicht.[9] Die Erstellung der *Datenbank Gesprochenes Deutsch* (DGD) wurde ab 1997 durch die Volkswagenstiftung finanziell gefördert und erlaubte letztendlich die komplette Überführung der digitalisierten Tonaufnahmen des Deutschen Spracharchivs in eine virtuelle, über das Internet zugängliche Version (Fiehler & Wagener 2005). Die DGD besteht aus unterschiedlichen Korpora, die sich wiederum aus unterschiedlichen Materialtypen zusammensetzen: Tonaufnahmen in unterschiedlichen Formaten (WAVE, WMA, MP3), Transkripte, die überwiegend als Fließtexte vorliegen, und Metadaten, die allgemeine Informationen über die Korpora und je nach Korpus ausführliche Informationen über die Sprecher der Interaktion und zur Situation der Aufnahme enthalten (Wagener 2005).

Mit der Digitalisierung und Archivierung analoger Tonaufnahmen seit den 1990er Jahren sowie dem Bereitstellen der Daten online folgt das IDS einem allgemeinen internationalen Trend zur digitalen Sprachdokumentation und archivierung. In Europa wurde z. B. im Jahr 2000 durch Finanzierung der Volkswagenstiftung das Projekt *Dokumentation bedrohter Sprachen* (DoBeS)[10] am Max-Planck-Institut (MPI) für Psycholinguistik in Nijmegen gegründet (Haig et al. 2012). Dieses Projekt setzt sich u. a. zum Ziel, vom Aussterben bedrohte Sprachen zu dokumentieren und die Ton- und Bildaufnahmen sowie die dazugehörigen Begleitmaterialien zu archivieren, um so Daten zu bedrohten Sprachen zu retten, und, soweit möglich, über das Internet für Forschungszwecke zur Verfügung zu stellen. So entstand im Rahmen von DoBeS in den letzten 17 Jahren nicht nur eine beträchtliche Anzahl von Sprachdokumentationsprojekten, sondern es wurde auch Software zur Aufnahme, Verarbeitung und Archivierung von Sprachaufnahmen entwickelt, die kostenfrei zur Verfügung gestellt wurde. Die so gesammelten wissenschaftlichen und technischen Erfahrungen wurden vor einigen Jahren am MPI Nijmegen in einer eigenen Abteilung (*The Language Archive* [TLA]) gebündelt, um so eine langfristige Unterstützung von Sprachdokumentations- und Archivierungsaktivitäten zu

9 Die Digitalisierung hat nicht nur die Archivierung und die Nutzung vorhandener Aufnahmen maßgeblich beeinflusst, sondern sie hat auch die linguistische Datenerhebung und -verarbeitung stark verändert. Siehe z. B. Bird & Simons (2003), Boas (2006) und Margetts & Margetts (2012).

10 http://dobes.mpi.nl (letzter Zugriff: 20. 6. 2017).

garantieren. Das am MPI angesiedelte *The Language Archive* ist genauso wie das IDS Mitglied im CLARIN-Verbund,[11] einer europäischen Forschungsinfrastruktur für Sprachressourcen und Sprachtechnologie.[12]

Gegenwärtig sind in der DGD mehrere Korpora aus Gebieten zugänglich, die als Sprachinseln des Deutschen gelten können. Umfangreich ist das Korpus zum Australiendeutsch, dessen Aufnahmen in den Jahren 1966 bis 1973 unter Leitung von Michael Clyne an der Monash University in Melbourne entstanden.[13] Ein weiterer umfangreicher Bestand liegt von Sprechern der ersten und zweiten Auswanderergeneration aus Israel vor, die im Wesentlichen in den 1990er und 2000er Jahren aufgenommen wurden. Ebenfalls existiert ein Korpus der ehemaligen deutschen Ostgebiete inklusive der Sprachinseln in Ost- und Südosteuropa, die zwischen 1962 und 1965 aufgenommen wurden, wobei dies überwiegend in Deutschland geschah. Im hauptsächlich zwischen 1955 und 1961 erhobenen Zwirner-Korpus, das den Versuch einer umfassenden Dokumentation der deutschen Dialekte darstellt, sind weitere Aufnahmen von Flüchtlingen, Vertriebenen und Übersiedlern aus diesen Gebieten gesammelt. Zuletzt findet sich im *Korpus Binnen- und Auslandsdeutsche Mundarten: Varia* eine Vielzahl von Aufnahmen, die in den 1960er und 1970er Jahren gemacht wurden. Neben den Aufnahmen aus dem deutschen Kerngebiet sind hier Sprecher aus Australien, Kanada, Mexiko und den USA dokumentiert.

Hier scheint es angebracht, einige Zahlen aufzuführen, um zum einen den Umfang dieser Korpora zu unterstreichen, gleichzeitig aber auch auf die Lücken hinzuweisen, die noch bestehen. Das Korpus *Australiendeutsch* kann mit 220 Aufnahmen mit einer Länge von mehr als insgesamt 64 Stunden als durchaus umfangreich gelten. Diese stammen aus 38 verschiedenen Orten in den Bundesstaaten South Australia und Victoria, wobei die Zahl der Aufnahmen pro Ort zwischen einer einzigen und 29 schwanken. Für 168 dieser 220 Aufnahmen ist in der DGD auch ein Transkript vorhanden. Das Korpus *Binnen- und Auslandsdeutsche Mundarten: Varia* ist dagegen deutlich stärker gestreut. Insgesamt enthält es zur Zeit 183 Aufnahmen mit mehr als 69 Stunden Länge. Von diesen stammen 3 aus Australien (sämtlich South Australia), 50 aus Deutschland, 26 aus Kanada (British Columbia 24, Ontario 2), 4 aus

---

11 CLARIN – *Common Language Resources and Infrastructure*, http://www.clarin.eu (letzter Zugriff: 20. 6. 2017).

12 Ähnliche Konsortien, die sich vorwiegend in Nordamerika mit dem Aufbau einer technischen-linguistischen Infrastruktur beschäftigen sind E-MELD – *Electronic Metastructure for Endangered Languages Data* (http://emeld.org [letzter Zugriff: 20. 6. 2017]) und OLAC – *Open Languages Archives Community* (http://www.language-archives.org/ [letzter Zugriff: 20. 6. 2017]).

13 Hier ist erwähnenswert, dass das IDS nicht sämtliche an der Monash University entstandenen Aufnahmen erhalten hat.

Mexiko (sämtlich Chihuahua), 12 aus Österreich, 12 aus der Schweiz sowie weitere 76 aus den USA (Illinois 9, Wisconsin 67). Insgesamt sind dies also 109 Aufnahmen aus deutschen Sprachinseln, Transkripte sind hier jedoch nicht vorhanden. Von diesen hat die überwiegende Mehrheit der Sprecher primär Englisch als Kontaktsprache, innerhalb dieser Gruppe sind jedoch mehr als 80 % aus nur zwei Regionen. Bezieht man das Korpus *Australiendeutsch* mit ein, wächst dieses Übergewicht der in einen englischen Kontext eingebetteten Sprachinseln zusätzlich. Für sprachwissenschaftliche Arbeit, die auf genau dieses Kontaktszenario abzielt, bietet die DGD damit eine wichtige Grundlage für empirische Arbeit. Sie ist jedoch räumlich eingeschränkt und wird auf sich allein gestellt einer vergleichenden Sprachinselforschung, wie sie später in diesem Beitrag diskutiert wird, noch nicht gerecht.

Neben diesen Aufnahmen besitzt das IDS weitere Daten, die jedoch bislang nicht öffentlich zugänglich sind. Dies sind etwa die Daten des Siebenbürgisch-Sächsischen Schallarchivs, welche jedoch auch über das Projekt Audioatlas Siebenbürgisch-Sächsischer Dialekte (ASD) an der Ludwig-Maximilians-Universität München verfügbar sind (siehe Abschnitt 3.2). Vorhanden aber nicht zugänglich sind dagegen Aufnahmen aus Panambi im brasilianischen Bundesstaat Rio Grande do Sul (vgl. Bärnert-Fürst 1994). Sowohl an diesen Daten aus Brasilien als auch an den von in die DGD integrierten Daten aus Australien lassen sich wichtige rechtliche Aspekte illustrieren, da diese die Nutzbarkeit der Daten grundlegend beeinflussen (mehr Details dazu in Gorisch, Schmidt & Stift i. V.). Für das IDS ist das Bundesdatenschutzgesetz (BDGS) verbindlich. Unter § 40 (3) heißt es:

> Die wissenschaftliche Forschung betreibenden Stellen dürfen personenbezogene Daten nur veröffentlichen, wenn
> 1. der Betroffene eingewilligt hat oder
> 2. dies für die Darstellung von Forschungsergebnissen über Ereignisse der Zeitgeschichte unerlässlich ist.[14]

Ob die Veröffentlichung von Aufnahmen aus deutschen Sprachinseln im Sinne von Satz 2 für die Darstellung von Forschungsergebnissen über Ereignisse der Zeitgeschichte unerlässlich ist, ist eine Frage, die sich nicht ohne weiteres bejahen lässt. Eindeutiger ist dagegen die Einwilligung der Sprecher. Im Falle der Daten aus Australien wurde damals das Einverständnis der Sprecher mündlich eingeholt und dies in der Regel aufgenommen. Neben diesem Regelfall gibt es jedoch auch Ausnahmen, in denen dieses Einverständnis nicht auf den Aufnahmen erscheint. Für die Veröffentlichung dieser Aufnahmen besteht damit

14 http://www.gesetze-im-internet.de/bdsg_1990/__40.html (letzter Zugriff: 19. 5. 2017).

keine hundertprozentige rechtliche Sicherheit, so dass im IDS davon abgesehen wird. In verwandter (doch anderer) Weise ist Datenschutz auch für die Daten aus Panambi ein Grund, der der Veröffentlichung bislang im Wege stand. In das aufgenommene Deutsch der Sprecher mischt sich auch oft Portugiesisch, welches in gleicher Weise anonymisiert werden muss wie die in Deutsch aufgenommenen Teile des Gesprächs. Mit entsprechenden Kapazitäten ist das IDS jedoch derzeit nicht ausgestattet, so dass andere Sprachinseldaten in der Priorität vor diese getreten sind.

Dieser kurze Einblick in den Stand der Archivierung und Digitalisierung von Sprachinselaufnahmen zeigt, dass das IDS eine zentrale Rolle in der Verortung und Verfügbarmachung entsprechender Sprachinselkorpora innehat. Er zeigt aber auch gleichzeitig, dass die Archivierung und Bereitstellung von Sprachinselaufnahmen im Internet auch in Zukunft weiterer Arbeit bedarf. Dass seit der Umfrage von Wagener & Bausch (1997) die Aufnahmen der Varietäten an bereits dokumentierten Orten fortgeführt oder an weiteren Orten neu begonnen wurden, erhärtet letzteren Befund. Gerade bei den Dokumentationsprojekten der letzten Jahre sind das IDS und seine Mitarbeiter jedoch sehr involviert. So leisten sie wichtige Unterstützung bei der Planung, Konzeption und Realisierung von digitalen Korpora, die von Forschungsprojekten auf der Basis aktueller linguistischer Feldforschungen zusammengestellt und später der DGD zugeführt werden. Besonders sei hier die zentrale Rolle von Thomas Schmidt genannt, der in den letzten Jahren in seiner Rolle als Leiter des Programmbereichs „Mündliche Korpora" am IDS nicht nur den Ausbau und die technische Modernisierung der DGD vorangetrieben hat, sondern auch gleichzeitig mehrere Forschungsprojekte wie Unserdeutsch (Peter Maitz, Universität Augsburg), und Namibiadeutsch (Horst Simon, FU Berlin, und Heike Wiese, Universität Potsdam) bei der Planung und Zusammenstellung ihrer digitalen Korpora technisch beraten und unterstützt hat.

3.2 Digitale Sprachinselkorpora außerhalb des IDS

Auch an anderen Institutionen gibt es digitale Sprachinselkorpora, wobei sich grundsätzlich drei Kategorien unterscheiden lassen. In die erste Kategorie fallen digitale Sprachinselkorpora, die auf bereits vorhandenen analogen Aufnahmen beruhen, welche digitalisiert worden sind und primär vor Ort zugänglich sind.

Ein Beispiel ist das *North American German Dialect Archive*, das im Max Kade Institute for German-American Studies an der University of Wisconsin-Madison beheimatet ist. Dort sind tausende Stunden von Aufnahmen deutscher Einwandererdialekte aus den USA vorhanden, die seit den 1940ern auf-

genommen wurden.[15] Die dort vorhandenen Daten stammen teilweise von an der University of Wisconsin ansässigen Forschern wie etwa Lester W. J. („Smoky") Seifert oder Jürgen Eichhoff, teilweise jedoch auch von Forschern an anderen Institutionen, so etwa die von Glenn Gilbert in den 1960er Jahren an der University of Texas erhobenen Aufnahmen des Texasdeutschen. Diese sind zwar mittlerweile digitalisiert worden, bislang ist jedoch nur ein Bruchteil der Bestände online in kurzen Ausschnitten zugänglich, und auch Archivlisten sind nicht online, so dass der Zugang nur umständlich möglich ist.[16] Eine Überarbeitung des Portals ist gegenwärtig in Planung, im Zuge derer auch detailliertere Informationen zu den Archivalien verfügbar gemacht werden sollen. Zusätzlich ist vorgesehen, zumindest Teile der Bestände Forschern auch aus der Ferne zugänglich zu machen.

Die zweite Kategorie umfasst online verfügbare digitale Sprachinselkorpora, die aus einer spezifischen Sorte linguistischer Daten aus einem bestimmten Gebiet bestehen. Ein Beispiel ist der *Linguistic Atlas of Kansas German*,[17] der die Sprachvariation unterschiedlicher deutscher Dialekte in Kansas mit Hilfe von Wenker-Sätzen erfasst. Nutzer der Online-Version können über eine Landkarte von Kansas (siehe den Ausschnitt aus der Webseite mit der interaktiven anklickbaren Karte von Kansas in Abb. 7.1) auf Tonaufnahmen von Wenker-Sätzen aus elf Landkreisen in Kansas zugreifen, insgesamt handelt es sich um Aufnahmen mit 21 Sprechern. Die Aufnahmen können auf der Webseite des *Linguistic Atlas of Kansas German* mit Hilfe eines eingebetteten Audio-Players abgespielt werden.[18]

Ein weiteres Beispiel ist die digitalisierte Version von Gesprächen mit ca. 20 Dialektsprechern wolgadeutscher Herkunft in North Dakota.[19] Diese Gespräche wurden in den 1970er Jahren von Allen Spiker aufgenommen, in den letzten paar Jahren digitalisiert und dann an der North Dakota State University online gestellt. Insgesamt handelt es sich um ca. 12 Stunden digitaler Tonaufnahmen, die man entweder mit einem eingebetteten Audio-Player hören oder als Datei im MP3-Format herunterladen kann.

---

15 https://mki.wisc.edu/content/north-american-german-dialect-archive (letzter Zugriff: 26. 5. 2017).

16 http://csumc.wisc.edu/AmericanLanguages/german/germ_us.htm (letzter Zugriff: 27. 5. 2017).

17 http://www2.ku.edu/~germanic/LAKGD/Atlas_Intro.shtml (letzter Zugriff: 20. 6. 2017).

18 Nicht alle vorhandenen Aufnahmen des Kansasdeutschen sind online gestellt. Es gibt insgesamt 136 Aufnahmen (zwischen 15–45 Minuten lang), die z. T. noch auf Kassetten sind (Chris Johnson, pers. Mitt.).

19 https://library.ndsu.edu/ndsuarchives/allen-spiker-german-russian-dialect-tapes (letzter Zugriff: 20. 6. 2017).

LAKGD

LAKGD Map/Home
Database Search
Other Recordings

Recorded Speakers
Privacy Disclaimer
Speaker Attributes

Questionnaires
Wenker Sentences
LAKGD Questionnaire

Identifying Dialects
Kansas German Dialects
German Homeland Dialects
Volga German Dialects in Russia

History of Germans in Kansas
Daitsch, Däätsch, Düütsch, Dietsch
Carman Map of German Settlements in Kansas
Carruth Map of Foreign Settlements in Kansas
Volga German Source Villages to Kansas
Census Maps 1870-1900
German Language Use in Kansas 2000-2010

About Us/Contact Us

Linguistic Atlas of Kansas German Dialects

As one travels throughout the state of Kansas, one cannot help noticing numerous place names which might lead one to believe that one is, indeed, not in Kansas anymore: Humboldt in Allen County, Bremen in Marshall County, Stuttgart in Phillips County, Marienthal in Wichita County, Windthorst in Ford County, Olmitz in Barton County, Olpe in Lyons County, Bern in Nemaha County, and many others. Whether named for famous German researchers (Alexander and Wilhelm von Humboldt), German political leaders of the nineteenth century (Ludwig Windthorst), cities and towns in Germany (Bremen, Stuttgart and Olpe), the capital of Switzerland (Bern), a city in Moravia in the former Austrian Empire (Olmütz), or a German colony near the Volga River in the Russian Empire (Marienthal), each of these Kansas communities is a living testament to the massive influx of German-speaking settlers who found new homes in Kansas during the period from the mid-1850s to the 1880s, and continue to immigrate to Kansas at the beginning of the twenty-first century. (Keel, 2006)

Kansas Language Symposium - German

Professor William Keel presented a talk on Kansas German dialects at the Kansas Languages Symposium, hosted by Johnson County Community College on November 8, 2012.

See the YouTube video of the presentation (approximate length 1 hour).

Darker counties indicate posted recordings. Click on a county to listen to German speakers from that county.

Abb. 7.1: Linguistic Atlas of Kansas German Dialects. Karteninterface zur Abfrage von Tondaten.

Als drittes Beispiel lässt sich das Korpus russlanddeutscher Sprecher in Sibirien an der Universität Göteborg nennen.[20] Die bisher online verfügbaren Tonaufnahmen stammen von Gesprächen mit vier Sprecherinnen, wobei jedes Gespräch zwischen 60–90 Minuten dauert.[21] Zu den Tonaufnahmen gibt es außerdem Transkripte, die knapp 4.000 Sätze umfassen (siehe auch Andersen 2016).[22]

20 https://spraakbanken.gu.se/korp/?mode=siberian_german#?stats_reduce=word&cqp=%5B%5D (letzter Zugriff: 20. 6. 2017). Zum Russlanddeutschen siehe auch Behrend (2003).

21 An der Universität Krasnoyarsk, Russland, gibt es außerdem ein regionales Forschungszentrum zur Erforschung und Dokumentation der Sprache und Kultur der Deutschen in Sibirien. Dort gibt es auch einige Tonaufnahmen online (http://deu.kspu.ru/ru/about-us.html [letzter Zugriff: 23. 6. 2017]).

22 Momentan wird auch an der Universität Erfurt von Csaba Földes das Projekt „Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus“ aufgebaut (https://www.ungarndeutsch.de/information [letzter Zugriff: 20. 6. 2017]), welches auch bald Tonaufnahmen und Transkripte über das Internet anbieten wird (Földes 2016).

Als letztes Beispiel sei nochmals der an der Ludwig-Maximilians-Universität München angesiedelte Audioatlas Siebenbürgisch-Sächsischer Dialekte genannt, der auf digitalisierten Tonaufnahmen aus den 1960er und 1970er Jahren basiert (Krefeld, Lücke & Mages 2016).[23] Die ca. 360 Stunden umfassenden Tonaufnahmen bestehen aus Wenker-Sätzen, Erzählungen, Märchen und Liedern, zu vielen der Aufnahmen gibt es auch Transkripte. Der Audioatlas kann mit Hilfeunterschiedlicher Abfragemasken durchsucht werden (Datenbankabfrage, Karten, Vergleich von Wenker-Sätzen etc.).

Die dritte Kategorie digitaler Sprachinselkorpora umfasst solche Korpora, die das direkte Resultat von aktuellen Forschungsprojekten sind. Solche Sprachinselprojekte und -korpora haben sich zum Ziel gesetzt, aktuell erhobene Sprachinselaufnahmen im Rahmen der Forschungstätigkeit zeitnah zu bearbeiten, zu transkribieren und zu archivieren. Ein Beispiel ist das im Jahre 2001 an der University of Texas at Austin gegründete *Texas German Dialect Project* (TGDP) (Boas 2002; Boas et al. 2010), das sich zum Ziel gesetzt hat, die letzten der noch ca. 6.000 verbliebenen Sprecher des Texasdeutschen aufzunehmen und die Aufnahmen zu archivieren.[24] Die vom TGDP erhobenen Daten gliedern sich in drei Teile. Zum einen werden in einem Interview systematisch biographische Informationen über Person und Leben der Sprecher sowie ihren Gebrauch der deutschen Sprache und ihre Spracheinstellungen erhoben (siehe Boas 2005, Boas & Fingerhuth 2017). Die so erhobenen Sprecherdaten bilden die Grundlage für die im *Texas German Dialect Archive* archivierten Tonaufnahmen, d. h. jede Tonaufnahme kann durch eine einzigartige Identifikationsnummer nur einem bestimmten Satz an Metadaten zugeordnet werden (siehe Boas 2006 und Abb. 7.2). Die von Eikel (1954) und Gilbert (1972) benutzten Übersetzungslisten werden vom TGDP wieder verwendet, um so einen kontrollierten Datensatz zu erstellen und um auch gleichzeitig feststellen zu können, wie sich das Texasdeutsche in den letzten 50–60 Jahren verändert hat. Als dritten Ansatz gibt es 30–60 Minuten lange soziolinguistische Interviews mit den Sprechern des Texasdeutschen, die sich, je nach Lage und Interesse um unterschiedliche Themen drehen (Vorfahren, Wetter, Politik, persönliche Interessen etc.). Außerdem wird teilweise ein Gespräch in Abwesenheit der Forscher nur unter Sprechern von Texasdeutsch aufgenommen, etwa beim Verrichten von Alltagsarbeiten oder beim Kartenspiel. In Summe kann die Erhebung sämtlicher Daten eines einzigen Sprechers ohne weiteres drei Stunden in Anspruch nehmen. Dies bietet grundsätzlich eine breite Datengrundlage für sprachwissenschaftliche Forschung. In der Praxis erweist es sich jedoch teilweise als

23 http://www.asd.gwi.uni-muenchen.de/ (letzter Zugriff: 20. 6. 2017).
24 http://www.tgdp.org (letzter Zugriff: 20. 6. 2017).

Abb. 7.2: Das Texas German Dialect Archive. Ausschnitt aus der Webseite (screen shot), mit Anzeige der transkribierten Interviews.

Problem, da z. B. das Interesse der Sprecher an der Teilnahme schwindet, und dadurch nur ein Teil der Daten erhoben werden kann. Von 2001–2018 hat das TGDP mehr als 660 Sprecher des Texas-Deutschen aufgenommen. Der Plan ist bis 2030 die Grenze von tausend Sprechern zu überschreiten, falls es bis dahin noch genügend Sprecher geben sollte (siehe Boas 2009).

Die vom TGDP digital aufgenommenen Eikel- und Gilbert-Sprachdaten werden nach den Interviews segmentiert und im *Texas German Dialect Archive* (TGDA) archiviert. Die soziolinguistischen Interviews werden mit ELAN (Wittenburg et al. 2006) transkribiert,[25] ins Englische übersetzt und ebenfalls im TGDA archiviert (für Details siehe Boas 2003/2006). Nutzer des Archivs können auf alle drei Sorten von Daten zugreifen, entweder über eine Datenbankabfrage, über ein geographisches Suchinterface mit Karten, auf denen die Erhebungsorte verzeichnet sind, oder über die Wort- und Satzlisten der Gilbert- und Eikel-Fragebögen. Die im TGDA archivierten Daten (Tonaufnahmen, Transkripte, Metadaten) sind nach Anmeldung und Einverständniserklärung mit den Nutzungsbedingungen online abrufbar entweder als MP3 im eingebetteten

25 https://tla.mpi.nl/tools/tla-tools/elan/ (letzter Zugriff: 20. 6. 2017).

Audio Player zusammen mit den Transkripten im HTML-Format oder zum Herunterladen als WAVE- (Ton) und EAF- (Text) Format zur Analyse und weiteren Bearbeitung mit ELAN (für Details, siehe Boas et al. 2010). Von den insgesamt mehr als 1.000 Stunden Tonaufnahmen sind bisher nur ca. 100 Stunden transkribiert und übersetzt.

Die Konzeption und technische Realisierung des Arbeitsablaufs des TGDP sowie die daraus resultierende Archivierung und Bereitstellung von Sprachaufnahmen und begleitenden Materialien im Internet wurde in den vergangenen Jahren von anderen Forschungsprojekten aufgegriffen und als Grundlage für deren eigene Aufnahme- und Archivierungsaktivitäten genutzt. So haben das *Indiana German Dialect Project* (Karen Roesch, Indiana University Purdue University Indianapolis), das *Wisconsin Low German Dialect Project* (Ryan Dux, Bucknell University), und das *Barossa-Deutsch Projekt* (Claudia Riehl, Ludwig-Maximilians-Universität München) (Riehl 2012) die vom TGDP verwendeten Fragebögen benutzt, um selbst vergleichbare Daten zu erheben. Die von diesen Forschergruppen erhobenen Daten sind nicht nur für sich alleine genommen interessant, sondern sie erlauben es auch zu vergleichen, was passiert, wenn unterschiedliche deutsche Auswandererdialekte in Kontakt mit unterschiedlichen Varietäten des Englischen stehen. Wie der folgende Abschnitt zeigt, werden momentan die technischen Voraussetzungen dafür geschaffen, dass Daten in digitalen Sprachinselkorpora systematisch miteinander verglichen werden können.

4 Aufgaben für die Zukunft: Lokalisierung, Digitalisierung und eine digitale vergleichende Sprachinselforschung

Die Sprachinselforschung hat sich im 20. Jahrhundert weitgehend dezentral entwickelt – ein Umstand, der sich vielleicht aus der Verteilung ihres Untersuchungsgegenstands und auch der Forschenden über sämtliche Kontinente erklären lässt. In dieser Diaspora ist daher eine Vielzahl von Korpora entstanden, zu deren Umfang und Status sich verlässliche Angaben nur teilweise machen lassen. Das oben besprochene Korpus Australiendeutsch in der DGD zeigt einen Fall, in dem historische Sprachinseltonaufnahmen digitalisiert und der Öffentlichkeit über das Internet zugänglich gemacht wurden. Die Digitalisierung und Aufnahme der Australiendeutschdaten in die DGD ist ein Glücksfall für die Forschung, da diese Daten sonst wahrscheinlich nicht langfristig gesichert gewesen wären, geschweige denn einem breiteren Forschungspubli-

kum zur Verfügung gestanden hätten. Doch leider ist dieses Vorgehen eher die Ausnahme.

In vielen Fällen scheinen die Sprachinseldaten (Tonaufnahmen, Transkripte, Begleitmaterialien etc.) weiterhin ausschließlich am Ort der ursprünglichen Erhebung oder bei der Person, die sie erhoben hat, vorzuliegen. Der Umfang dieser Daten dürfte beachtlich sein, und sie sind, wenn überhaupt, nur vor Ort abzufragen, wobei es gerade bei älteren Magnetbändern nicht klar ist, wie lang diese noch haltbar sind. Auf Grundlage der von Wagener und Bausch durchgeführten Befragung lassen sich als Beispiele hierfür etwa unter Leitung von Evelyn S. Firchow in Minnesota geführte Interviews nennen, wo zum Zeitpunkt der Erhebung Anfang der 1990er Jahre bereits ca. 150 Interviews vorlagen, jedoch auch weitere Aufnahmen geplant waren (Wagener & Bausch 1997: 223, vgl. auch Firchow 1991). Im rumänischen *Arhiva de Folclor a Academiei Române* (sowie als Kopie an der siebenbürgisch-säschsischen Forschungsstelle in Gundelsheim – seit 2003 als Siebenbürgen-Institut ein An-Institut der Universität Heidelberg – dort jedoch nicht zugänglich) liegen nach Wagener & Bausch (1997: 208) 85 Magnetophonbänder mit Aufnahmen aus Siebenbürgen. Es ließen sich weitere Beispiele anführen, doch sollten allein diese beiden genügen, um zu illustrieren, dass eine beträchtliche Menge von Sprachdaten existiert, jedoch derzeit nicht oder nur schwer zugänglich ist. Ferner ist unsicher, ob und wie eine Archivierung und Sicherung der auf analogen Datenträgern erhobenen Daten in den unterschiedlichen Institutionen durchgeführt wird, geschweige denn ob Pläne vorhanden sind, diese einer breiteren Öffentlichkeit über das Internet zur Verfügung zu stellen.

Hier ist zu erwähnen, dass die von Wagener und Bausch durchgeführte Erhebung, die wohl als letzte und einzige ihrer Art gelten kann, einige in diesem Artikel bereits erwähnte Bestände nicht erfasst. Die Dokumentation der existierenden Korpora ist also lückenhaft. Ferner gibt es Aufnahmen, die derzeit als verschollen gelten müssen, wie etwa Eikels Aufnahmen des Texasdeutschen aus den 1940er Jahren. Eine entsprechende Übersicht existierender Sprachinseltonaufnahmen, die von der Dokumentation durch Wagener und Bausch ausgeht, findet sich im Anhang.

Für die Gegenwart stellen sich also die Aufgaben, die existierenden Aufnahmen zu erfassen, zu lokalisieren, zu digitalisieren, zu archivieren, sie für die Nutzung aufzubereiten und sie letztlich der Fachöffentlichkeit über das Internet zur Verfügung zu stellen. Dadurch ließen sich nicht nur die Früchte vergangener Arbeit sichern. Vielmehr bietet sich dadurch die Möglichkeit, der Sprachinselforschung die Daten und Infrastruktur zur Verfügung zu stellen, die die Erschließung eines Forschungsparadigmas ermöglicht, das bereits von Rosenberg (2003, 2005) angedacht wurde, das aber bislang noch nicht fest

etabliert ist: einer vergleichenden Sprachinselforschung. Boas (2016) skizziert, wie eine vergleichende Sprachinseldatenbank aussehen könnte. Konkret wird vorgeschlagen,

> [...] auf der Basis des TGDA eine erweiterte Sprachinseldatenbank aufzubauen, in der Daten von anderen Sprachinseln parallel korpuslinguistisch verarbeitet und archiviert werden. Der Umfang, die Qualität und die Art von bereits existierenden Sprachinseldaten sind natürlich recht unterschiedlich. [...] Aber alle bereits existierenden Daten könnten prinzipiell mit denselben im Rahmen des TGDP erprobten sowie evtl. noch zu erforschenden zusätzlichen Methoden korpuslinguistisch verarbeitet werden, um so in separaten Sprachinselarchiven archiviert zu werden. Diese Sprachinselarchive könnten dann miteinander vernetzt werden, um eine vergleichende Onlinesprachinseldatenbank zu implementieren. Ziel einer solchen vergleichenden Sprachinseldatenbank ist nicht nur die Archivierung existierender Sprachinseldaten für künftige Generationen, sondern auch die Bereitstellung der Sprachinseldaten für die vergleichende Sprachinselforschung. (Boas 2016: 39)

Die im Jahr 2016 durchgeführte Erneuerung der technischen Infrastruktur des TGDA an der University of Texas at Austin hat dazu geführt, dass die zugrunde liegende Datenbank nun auch Daten aus anderen Sprachinseln aufnehmen kann. Bei den Sprachinseldokumentationsprojekten, die sich an das TGDP anlehnen und vergleichbare Fragebögen und Einwilligungserklärungen benutzen, ist die Aufnahme von neuen Sprachinseldaten relativ einfach. Die ersten Schritte, Sprachinseldaten aus Wisconsin, Indiana und Australien in das vergleichende Sprachinselarchiv aufzunehmen, waren erfolgreich. Diese werden im Laufe des Jahres 2019 auch für die Öffentlichkeit freigeschaltet.

Etwas anders sieht die Lage bei älteren Sprachinselaufnahmen aus, bei denen es keine Einwilligungserklärungen gibt. Die Probleme bei der Zugänglichmachung von Daten in Deutschland wurden bereits im Kontext der über das IDS verfügbaren Korpora diskutiert. Aufgrund der unterschiedlichen Gesetzeslage gestaltet sich die Praxis in den USA anders. Forschung mit menschlichen Teilnehmern muss grundsätzlich von einem sogenannten *Institutional Review Board* (IRB) genehmigt werden.[26] Diese Genehmigung hängt neben grundsätzlichen ethischen Kriterien nicht zuletzt vom Einverständnis der Teilnehmer ab. Diese Genehmigungspflicht hat sich jedoch erst seit den 1970ern entwickelt. Abseits der Forschung sind jedoch persönliche Daten als solche in den USA im Gegensatz zu etwa Deutschland nicht in gleicher Weise geschützt. Anwenden lassen sich jedoch die Regularien des sogenannten *Fair Use* aus

26 Details zu den Begriffen „Forschung“ und „menschlicher Teilnehmer“ finden sich etwa unter https://research.utexas.edu/ors/human-subjects/what-is-human-subjects-research/ (letzter Zugriff: 18. 6. 2017).

dem Bereich des Urheberrechts. Nach diesen lassen sich jegliche Materialien für Zwecke wie Forschung oder Bildung verwenden, wenn Umfang und Art der Nutzung dies rechtfertigen.

Die Kriterien für eine solche Nutzung sind nicht rigide definiert, sondern verlangen eine Abschätzung von Seiten des Nutzers, und die Verwendung alter Aufnahmen für Forschungszwecke lässt sich vor diesem Hintergrund rechtfertigen. Eine Verpflichtung zur Wahrung der Persönlichkeitsrechte der Sprecher, wie sie bei der Bewilligung neuer Forschung durch das IRB gefordert würde, bleibt dabei jedoch weiterhin bestehen. Auch alte Sprachaufnahmen können deshalb nicht ohne weiteres veröffentlicht werden, sondern müssen soweit wie möglich anonymisiert werden. Neben der Digitalisierung analoger Sprachinselaufnahmen sowie der Anfertigung von Transkripten stellt die Anonymisierung daher einen recht zeitaufwendigen Arbeitsschritt dar. Im ersten Halbjahr 2017 wurde die erste Testphase zur Erschließung und Einbindung bereits existierender historischer Sprachinselaufnahmen aus Pennsylvania und Brasilien erfolgreich abgeschlossen. Diese Daten werden auch im Verlauf des Jahres 2019 der Öffentlichkeit im vergleichenden Sprachinselarchiv der University of Texas at Austin zur Verfügung gestellt.

Die Zusammenführung unterschiedlicher Sprachinselkorpora unter einem Dach hat den Vorteil, dass man viele linguistische Phänomene besser und intensiver untersuchen kann,

> […] da eine vergleichende Sprachinseldatenbank es erlaubt, nach Daten zu bestimmten sprachlichen Phänomenen zu suchen und diese dann systematisch zwischen Sprachinseln zu vergleichen. So könnte der Nutzer einer vergleichenden Sprachinseldatenbank z. B. nach bestimmten Präpositionen suchen, um zu sehen, welche Kasus diese in unterschiedlichen Sprachinselvarietäten markieren. Ein vergleichendes Sprachinselarchiv hätte nicht nur den Vorteil, dass bereits existierende Daten gesichert und archiviert würden und so der Nachwelt erhalten blieben, sondern es würde endlich auch eine wirklich systematische vergleichende Sprachinselforschung ermöglichen. Forscher, die ihre Sprachinseldaten in einem solchen Archiv deponieren würden, erhielten außerdem die Möglichkeit, die Früchte ihrer langjährigen Arbeit zu teilen. Entsprechende Richtlinien stellen außerdem sicher, dass die Forscher, die ihre Daten anderen Kollegen zur Verfügung gestellt haben, durch entsprechende Zitate ihrer Daten Anerkennung erhalten. (Boas 2016: 39–40)

Mit diesem Schritt ließe sich die Sprachinselforschung in ihrer Breite sinnbildlich aus dem analogen ins digitale Zeitalter, aus dem 20. Jahrhundert in die Gegenwart überführen und die Grundlage für zukünftige Arbeit schaffen. Während die bisherige Forschung mit ihrem Fokus auf einzelne Varietäten eine unerlässliche Grundlage gelegt hat, bietet der vergleichende Ansatz neue Fragen und Erkenntnisse, die auf Muster jenseits der einzelnen Varietäten oder Kontaktsprachen hinzielen. Weiter könnte dazu beigetragen werden,

das Potenzial der Sprachinselforschung als Schnittstelle zwischen der Sprachwissenschaft des Deutschen und Nachbardisziplinen zu verwirklichen. Sprachinselforschung ist fast notwendig Sprachkontaktforschung, die Perspektive der Kontaktsprache ist jedoch bislang weitgehend unbeachtet geblieben.

5 Zusammenfassung und Ausblick

Das Gesagte hinterlässt ein durchwachsenes Bild vom gegenwärtigen Stand der Sprachinselkorpora des Deutschen, erlaubt aber gleichzeitig einen positiven Ausblick. Angelegt in ihrem zergliederten Forschungsgegenstand ist die Sprachinselforschung weitgehend unkoordiniert entstanden. Die erhobenen Sprachkorpora sind diesem Umstand entsprechend ebenfalls zerstreut. Um das Potenzial der in einem Zeitraum von mehr als einem Jahrhundert gesammelten Daten zu nutzen, stellen sich in der Gegenwart die Aufgaben, die existierenden Aufnahmen zu lokalisieren, zu digitalisieren, und innerhalb der gesetzlich gesteckten Rahmen zugänglich zu machen.[27]

Es gibt jedoch bereits Ansätze zur Lösung dieser Aufgabe. Dem IDS kommt dabei eine zentrale Rolle zu, und die Übernahme der von Michael Clyne in Australien gemachten Aufnahmen sowie deren Veröffentlichung in der DGD können als ein Beispiel dafür angeführt werden, wie ein analog erhobenes Korpus von beachtlichem Umfang in ein digitales und über das Internet weltweit zugängliches umgewandelt wurde. Die Bedeutung des IDS wächst zusätzlich durch seine Betreuung bei der Planung von neuen Erhebungen, wie etwa im Fall der gegenwärtigen Arbeit am Namibiadeutschen oder am Unserdeutsch in Papua Neu Guinea und Australien.

Außerhalb des IDS gibt es ferner mit dem vergleichenden Sprachinselarchiv an der University of Texas at Austin Bemühungen, eine Infrastruktur aufzubauen, die gezielt auf die Bereitstellung von Sprachinselkorpora hin ausgerichtet ist. Dieses Sprachinselarchiv besitzt das Potenzial, die Sprachinselforschung dabei zu unterstützen, über ihren bisherigen Forschungsansatz hinauszuwachsen, der im Wesentlichen auf einzelne Varietäten beschränkt ist. Auf Grundlage zentralisiert zugänglicher Korpora könnte so eine systematisch

27 Ohne vom Gegenstand der Korpora des gesprochenen Deutschen abweichen zu wollen, scheint es doch angemessen, an dieser Stelle auf Lücken in der Erforschung des geschriebenen Deutsch der Sprachinseln hinzuweisen. Dieses ist in der Forschung bisher nur vereinzelt genutzt worden, etwa in Form von Auswandererbriefen (etwa Elspaß 2005), aber auch in der Sprachinselpresse (Földes 2015). Das weitere Erschließen solcher schriftsprachlicher Ressourcen, gegenwärtig wie auch historisch, stellt eine weitere Aufgabe für die Sprachinselforschung dar.

vergleichende Sprachinselforschung entstehen, die auf Grundlage historischer und neu erhobener Korpora mit neuen Fragen Forschung betreibt. Sie könnten Brücken zu anderen Disziplinen schlagen und zur Verankerung der Sprachinselforschung jenseits ihrer traditionellen Heimat in der Germanistischen Linguistik beitragen.

Literatur

Andersen, Christiane (2016): Nachfeld im Kontakt. Eine Korpusuntersuchung am Russlanddeutschen in Sibirien. *Göteborger Arbeitspapiere zur Sprachwissenschaft* 6, 1–15.

Atwood, E. Bagby (1962): *The regional vocabulary of Texas*. Austin, TX: University of Texas Press.

Bärnert-Fürst, Ute (1994): Conversation and displacement processes of the German language in the speech community of Panambi, Rio Grande Do Sul, Brazil. In Nina Berend & Klaus J. Mattheier (Hrsg.), *Sprachinselforschung. Eine Gedenkschrift für Hugo Jedig*, 273–287. Frankfurt am Main: Peter Lang.

Berend, Nina (2003): Zur Vergleichbarkeit von Sprachkontakten: Erfahrungen aus wolgadeutschen Sprachinseln in den USA und Russland. In William D. Keel & Klaus J. Mattheier (Hrsg.), *German language varieties worldwide: Internal and external perspectives (Deutsche Sprachinseln weltweit: Interne und externe Perspektiven)*, 239–269. Frankfurt am Main: Peter Lang.

Bird, Steven & Gary Simons (2003): Seven dimensions of portability for language documentation and description. *Language* 79 (4), 557–582.

Boas, Hans C. (2002): The Texas German Dialect Archive as a tool for analyzing sound change. In Peter Austin, Helen A. Dry & Peter Wittenburg (Hrsg.), *Proceedings of the International Workshop on Resources and Tools in Field Linguistics held in conjunction with the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain.

Boas, Hans C. (2003): Tracing dialect death: The Texas German Dialect Project. In Julie Larson & Mary Paster (Hrsg.), *Proceedings of the 28th Meeting of the Berkeley Linguistics Society*, 387–398. Berkeley, CA: Berkeley Linguistics Society.

Boas Hans C. (2005): A dialect in search of its place: The use of Texas German in the public domain. In Craig Cravens & David Zersen (Hrsg.), *Transcontinental encounters: Central Europe meets the American heartland*, 78–102. Austin, TX: Concordia University Press.

Boas, Hans C. (2006): From the field to the web: Implementing best-practice recommendations in documentary linguistics. *Language Resources and Evaluation* 40 (2), 153–174.

Boas, Hans C. (2009): *The life and death of Texas German*. Durham: Duke University Press.

Boas, Hans C. (2016): Variation im Texasdeutschen: Implikationen für eine vergleichende Sprachinselforschung. In Alexandra Lenz (Hrsg.), *German Abroad. Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung*, 11–44. Wien: Vienna University Press.

Boas, Hans C. & Matthias Fingerhuth (2017): „I am proud of my language but I speak it less and less!" – Der Einfluss von Spracheinstellungen und Sprachgebrauch auf den Spracherhalt von Heritage-Sprechern des Texasdeutschen. *Linguistische Berichte* 249, 95–121.

Boas, Hans, C., Marc Pierce, Karen Roesch, Guido Halder & Hunter Weilbacher (2010): The Texas German Dialect Archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics* 22 (3), 277–296.

Bremer, Otto (1895): *Beiträge zur Geographie der deutschen Mundarten in Form einer Kritik von Wenkers Sprachatlas des Deutschen Reichs*. Leipzig. Breitkopf & Härtl.

Brenner, Koloman, Maria Erb & Karl Manherz (2008): *Ungarndeutscher Sprachatlas*. Band 1. Budapest: ELTE Germanistisches Institut.

Eikel, Fred (1949): The use of cases in New Braunfels German. *American Speech* 24 (4), 278–281.

Eikel, Fred (1954): *The New Braunfels German dialect*. Manuskript. Johns Hopkins University.

Eikel, Fred (1966): New Braunfels German: Part I. *American Speech* 41 (1), 5–16.

Eller-Wildfeuer, Nicole (im Druck): *Sprecherbiographien und Mehrsprachigkeit. Deutschbasierte Minderheitensprachen in Osteuropa und Übersee*. Tübingen: Stauffenburg.

Elspaß, Stephan (2005): *Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert*. Tübingen: Niemeyer.

Fiehler, Reinhard & Peter Wagener (2005): Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Dokumentation und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 6, 136–147.

Firchow, Evelyn Scherabon (1991): Deutsche Sprachinseln im amerikanischen Bundesstaat Minnesota. In Yoshinori Shichiji (Hrsg.), *Begegnung mit dem „Fremden". Grenzen, Traditionen, Vergleiche; Akten des VIII. Internationalen Germanisten-Kongresses, Tokyo 1990. Band 3: Sprachgeschichte: Sprachkontakte im germanischen Sprachraum*, 252–263. München: Iudicium.

Földes, Csaba (2015): Literalität im Schnittfeld von zwei Sprachen und Kulturen. Beobachtungen anhand der Phraseologie in der Sprache der Lokalpresse. In Regula Schmidlin, Heike Behrens & Hans Bickel (Hrsg.), *Sprachgebrauch und Sprachbewusstsein. Implikationen für die Sprachtheorie*, 239–260. Berlin, Boston: de Gruyter.

Földes, Csaba (2016): Ungarndeutsche Sprachvariation und Mehrsprachigkeit. Ein Korpusprojekt auf der Basis empirischer Feldforschung und Online-Sprachdokumentation. *Sprachtheorie und Germanistische Linguistik* 26 (2), 167–190.

Gilbert, Glenn G. (1963): *The German dialect spoken in Kendall and Gillespie Counties, Texas*. PhD Dissertation, Harvard University.

Gilbert, Glenn G. (1972): *The Linguistic Atlas of Texas German*. Austin, TX: The University of Texas Press.

Goebl, Hans & Guillaume Schiltz (2006): Neuere Entwicklungen in der europäischen Dialektologie (1950–2000). In Sylvain Auroux, E. F. K. Koerner, Hans-Josef Niederehe & Kees Versteegh (Hrsg.), *Geschichte der Sprachwissenschaften. Ein internationales Handbuch zur Entwicklung der Sprachforschung von den Anfängen bis zur Gegenwart*, 3. Teilband (Handbücher zur Sprach- und Kommunikationswissenschaft 18.3), 2352–2362. Berlin, Boston: Walter de Gruyter.

Gorisch, Jan, Schmidt, Thomas & Ulf Michael Stift (i.V.): Data of German speech minorities in the Archive for Spoken German: An overview. Manuskript. IDS Mannheim.

Haig, Geoffrey, Nicole Nau, Stefan Schnell & Claudia Wegener (2012): Introduction: Documenting endangered language before, during, and after the DoBes programme. In Geoffrey Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (Hrsg.), *Documenting*

endangered languages: Achievements and perspectives, 1–14. Berlin, Boston: de Gruyter.

Herrgen, Joachim (2001): Die Rolle des Wenker-Atlasses in der Geschichte der Dialektologie. In Sylvain Auroux, E. F. K. Koerner, Hans-Josef Niederehe & Kees Versteegh (Hrsg.), *Geschichte der Sprachwissenschaften. Ein internationales Handbuch zur Entwicklung der Sprachforschung von den Anfängen bis zur Gegenwart*, 2. Teilband (Handbücher zur Sprach und Kommunikationswissenschaft 18.2), 1513–1535. Berlin, Boston: de Gruyter.

Herrgen, Joachim & Alexandra Lenz (2003): Digitale Dialektologie. Online Publikation des Wenker Atlasses im Internet. *Marburger UniJournal* 14, 43–48.

John, Eckhard & Natalia D. Swetosarowa (Hrsg.) (2005): *Traditionelle Lieder der Russlanddeutschen. Die Volksliedsammlung von Viktor M. Schirmunski. Ein Quellenhandbuch*. Münster: Waxmann.

Klein, Karl Kurt & Ludwig Erich Schmitt (Hrsg.) (1961/1965): *Siebenbürgisch-Deutscher Sprachatlas*. 2 Bände. Marburg: Elwert.

Knetschke, Edeltraud & Margret Sperlbaum (1983): *Das Deutsche Spracharchiv im Institut für Deutsche Sprache* (Mitteilungen des Instituts für Deutsche Sprache 6). 2. Aufl. Mannheim: Institut für Deutsche Sprache.

Knoop, Ulrich, Wolfgang Putschke & Herbert-Ernst Wiegand (1982): Die Marburger Schule: Entstehung und frühe Entwicklung der Dialektgeographie. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert-Ernst Wiegand (Hrsg.), *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. 1. Teilband (Handbücher zur Sprach- und Kommunikationswissenschaft 1.1), 38–92. Berlin, New York: Walter de Gruyter.

Krefeld, Thomas, Lücke, Stephan & Emma Magnes (Hrsg.) (2016): *Zwischen traditioneller Dialektologie und digitaler Geolinguistik: Der Audioatlas siebenbürgisch-sächsischer Dialekte (ASD)* (Korpus im Text 2). Münster: Monsenstein & Vannerdat.

Kuhn, Walter (1934): *Deutsche Sprachinselforschung. Geschichte, Aufgaben, Verfahren*. Plauen: Wolff.

Lessiak, Primus & Anton Pfalz (1918): Sprachproben aus den Sieben Gemeinden (Sette Comuni Vicentini), Italien. *Sitzungsberichte/Akademie der Wissenschaften in Wien, Philosophisch-Historische Klasse* 187 (1), 59–74.

Lüdeling, Anke & Merja Kytö (Hrsg.) (2008/2009): *Corpus linguistics. An international handbook*, 2 Bde. (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1–2). Berlin, Boston: Mouton de Gruyter.

Margetts, Anna & Andrew Margetts. 2012. Audio and video recording techniques for linguistic research. In Nick Thieberger (Hrsg.), *The Oxford handbook of linguistic fieldwork*, 13–53. Oxford: Oxford University Press.

Moelleken, Wolfgang Wilfried (1988): A new linguistic atlas of Pennsylvania German. *Monatshefte* 80 (1), 105–114.

Reed, Carroll & Lester W. Seifert (1954): *A linguistic atlas of Pennsylvania German*. Marburg: Becker.

Riehl, Claudia (2012): Deutsch als Reliktvarietät: Der Fall des Barossa-Deutschen (Australien). In Claudia Riehl & Elisabeth Knipf-Komlosi (Hrsg.), *Kontaktvarietäten des Deutschen in historischer und gegenwärtiger Sicht*, 37–40. Wien: edition praesens.

Rosenberg, Peter (2003): Comparative speech island research: Some results from studies in Russia and Brazil. In Willian D. Keel & Klaus J. Mattheier (Hrsg.), *German language varieties worldwide: Internal and external perspectives*, 199–238. Frankfurt am Main: Peter Lang.

Rosenberg, Peter (2005): Dialect convergence in the German language islands (*Sprachinseln*). In Peter Auer, Frans Hinskens & Paul Kerswill (Hrsg.), *Dialect change: Convergence and divergence in European languages*, 221–235. Cambridge: Cambridge University Press.

Schirmunski, Viktor M. (1962): *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Aus dem Russischen übersetzt und wissenschaftlich bearbeitet von Wolfgang Fleischer. Berlin: Akademie Verlag.

Schmeller, Johann Andreas (1821): *Die Mundarten Bayerns grammatisch dargestellt*. München: Karl Thienemann.

Schmeller, Johann A. (1838): Ueber die sogenannten Cimbern der VII und XIII Communen auf den Venedischen Alpen und ihre Sprache. *Abhandlungen der Bayerischen Akademie der Wissenschaften, Philosophisch-Philologische und Historische Klasse*, 559–708. http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb10523345-1 (letzter Zugriff: 23. 6. 2017).

Schüller, Dietrich (2008): *Audiovisual research collections and their preservation*. Amsterdam: European Comission on Preservation and Access.

Schweizer, Bruno (1939): *Zimbrische Sprachreste. Teil 1: Texte aus Giazza (Dreizehn Gemeinden ob Verona)*. Nach dem Volksmunde aufgenommen und mit deutscher Übersetzung herausgegeben. Halle (Saale): Niemeyer.

Stift, Ulf-Michael & Thomas Schmidt (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In Melanie Steine & Franz Josef Berens (Hrsg.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 360–375. Mannheim: Institut für Deutsche Sprache. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-24779 (letzter Zugriff 23. 6. 2017).

Wagener, Peter (1988): *Untersuchungen zur Methodologie und Methodik der Dialektologie*. Marburg: Elwert.

Wagener, Peter (2002): Gesprochenes Deutsch online. Zur Modernisierung des Deutschen Spracharchivs. *Zeitschrift für Dialektologie und Linguistik* 3, 314–335.

Wagener, Peter (2005): Die Datenbank Gesprochenes Deutsch. Archivierung, Dokumentation und Erschließung des Deutschen Spracharchivs (DSAv). *IDS Sprachreport* 3/2005, 23–26.

Wagener, Peter & Karl-Heinz Bausch (Hrsg.) (1997): *Tonaufnahmen des gesprochenen Deutsch. Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven*. Tübingen: Niemeyer.

Wildfeuer, Alfred (2017a): Sprachinseln, Sprachsiedlungen, Sprachminderheiten. Zur Bezeichnungsadäquatheit dieser und weiterer Termini. In Alexandra Lenz, Ludwig Breuer, Tim Kallenborn, Peter Ernst, Manfred Glauninger & Franz Patocka (Hrsg.), *Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts – Dynamik, Struktur, Funktion* (Zeitschrift für Dialektologie und Linguistik – Beihefte 167), 373–388. Stuttgart: Steiner.

Wildfeuer, Alfred (2017b): *Sprachenkontakt, Mehrsprachigkeit und Sprachverlust. Deutschböhmisch-bairische Minderheitensprachen in den USA und in Neuseeland* (Linguistik – Impulse & Tendenzen 73). Berlin, Boston: de Gruyter Mouton.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes (2006). ELAN: A professional framework for multimodality research. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, *Genoa, Italy*, 1556–1559. Paris: European Language Resources Association (ELRA).

Zwirner, Eberhard (1983): Fünfzig Jahre Deutsches Spracharchiv. *Zeitschrift für Dialektologie und Linguistik* 50 (1), 35–43.

Anhang: Sprachinselkorpora

Region/Land	Zeitraum	Hauptverantwortlich	Umfang	Wo vorhanden?[28]	Weitere Informationen
Colonia Tovar, Venezuela	1986/1987	Klaus Lückert	68 Aufnahmen	Privat?	Wagener & Bausch (1997: 4–6)
Israel	1989–1994, 2000–2011	Anne Betten, Kristine Hecker	178 Aufnahmen	über DGD zugänglich (Korpora IS, ISW, ISZ)	Wagener & Bausch (1997: 51–53)
Siebenbürgen, Rumänien	1966–1975	Ruth Kisch, Heinrich Mantsch	190 Aufnahmen	Ludwig-Maximilians-Universität München (Zugriff möglich), IDS (kein Zugriff)	Wagener & Bausch (1997: 69), http://www.asd.gwi.uni-muenchen.de/
Slowakei	1977–1980	R. Melzer	39 Tonbänder	Karpatendeutsches Heimatmuseum, Karlsruhe	Wagener & Bausch (1997: 85)
Ehemalige Sowjetunion, USA, Kanada, Mexiko, Paraguay	1959–1991	Ulrich Tolksdorf	Ungewiss, Teil einer Sammlung mit 544 Aufnahmen	Privatarchiv Tolksdorf, Kiel?	Wagener & Bausch (1997: 91–92)
Nord- und Südamerika	1980–?	Ulrich Tolksdorf	80 Aufnahmen	Universität Kiel? Evtl. Deutscher Sprachatlas Marburg	Wagener & Bausch (1997: 92–96)
Panambi, Brasilien	1985	Ute Bärnert-Fürst	23 Aufnahmen	Universität Mannheim, IDS	Wagener & Bausch (1997: 144–145)

28 Sofern die Autoren ermitteln konnten, dass es einen Wechsel in der die Aufnahmen beherbergenden Institution gegeben hat, ist dies hier aktualisiert. Dabei wird grundsätzlich davon ausgegangen, dass die neue Institution die Bestände weiterhin verwahrt. Wenn unklar ist, ob der bei Wagener & Bausch (1997) angegebene Verbleibeort aufgrund von Institutionswechseln oder ähnlichem weiterhin zutrifft, ist dies mit einem Fragezeichen markiert.

Region/Land	Zeitraum	Hauptverantwortlich	Umfang	Wo vorhanden?	Weitere Informationen
Ungarn, Slowakei	1957–1992	Alfred Camman	Ungewiss, Teil einer Sammlung von ca. 1.000 Aufnahmen	Verein der Freunde des Archivs für Heimatforschung in Rotenburg (Wümme) e. V.	Wagener & Bausch (1997: 179–180)
Südosteuropa	1954–1992	Hans Gehl (Überspielung)	212 Aufnahmen	Institut für donauschwäbische Landeskunde, Tübingen	Wagener & Bausch (1997: 192–194
Diverse USA	1964–?	Wolfgang W. Moelleken	ca. 430 Aufnahmen	University at Albany?, IDS (teilweise), Max-Kade-Institute	Wagener & Bausch (1997: 202–205)
Wisconsin, USA	1968–1969	Jürgen Eichhoff	64 Aufnahmen	Max-Kade-Institute, IDS (teilweise)	Wagener & Bausch (1997: 222–223)
Minnesota, USA	1986–?	Eveylin S. Firchow	ca. 150 Aufnahmen	University of Minnesota?	Wagener & Bausch (1997: 222–223)
Italien, Tschechien, Slowakei, Ungarn u. a.	1958–?	Maria Hornung	ca. 500 Aufnahmen	Verein der Freunde der von Österreich aus im MA besiedelten Sprachinseln/Verein der Sprachinselfreunde	Wagener & Bausch (1997: 247–248)
Brasilien, Neuseeland, Rumänien, Ukraine, USA (Kansas und Minnesota)	2005–2013	Nicole Eller-Wildfeuer, Alfred Wildfeuer	unter anderem ca. 50 Stunden Tonaufnahmen	Privat	Eller-Wildfeuer (im Druck) Wildfeuer 2017b

Christoph Draxler und Florian Schiel

8 Moderne phonetische Datenbanken

Erstellung und Datenaufbereitung

Abstract: Dieser Beitrag befasst sich mit Korpora und Datenbanken, die als empirische Grundlage phonetischer Analysen verwendet werden. Moderne phonetische Datenbanken stehen in einer starken Wechselwirkung zwischen Phonetik und Sprachtechnologie: die Sprachtechnologie liefert technische Verfahren zur Speicherung und Analyse gesprochener Sprache und ermöglicht damit überhaupt erst die Verarbeitung auch großer Mengen an Sprachdaten. Im Gegenzug sind Erkenntnisse der Phonetik die Basis vieler sprachtechnologischer Verfahren bzw. tragen zu ihrer Verbesserung bei. Der Beitrag gliedert sich in zwei Teile: im ersten Teil werden die Datenarten phonetischer und sprachtechnologischer Datenbanken beschrieben. Im zweiten Teil wird der in der Praxis relevante Prozess der Erstellung und Nutzung einer phonetischen Datenbank anhand eines konkreten Projekts präsentiert; der Fokus liegt hierbei auf den fachlichen Aspekten: Forschungsfrage, Datensammlung, -aufbereitung und -auswahl sowie Analyse und Interpretation.

Keywords: Datenbanken, Phonetik, Sprachtechnologie

1 Gegenstand und Motivation

Untersuchungsgegenstand der Phonetik ist nach Pompino-Marschall (1995: 3) der „lautliche Aspekt der sprachlichen Kommunikation“. Becker (2012) schreibt in Abgrenzung zur Phonologie:

> Die *Phonetik* […] beschreibt die materielle Seite der Laute sprachlicher Äußerungen, die Abläufe der Sprachproduktion und -wahrnehmung durch die Sprecher, einschließlich der kognitiven und neuronalen Aspekte, mit naturwissenschaftlichen Methoden, etwa mit Experimenten oder Messungen, ohne unmittelbare Berücksichtigung des Sprachsystems. (Becker 2012: 13)

Christoph Draxler, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: draxler@phonetik.uni-muenchen.de
Florian Schiel, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: schiel@phonetik.uni-muenchen.de

https://doi.org/10.1515/9783110538649-009

Reetz & Longman (2009) unterteilen die Phonetik in die drei traditionellen Teilbereiche Produktion, Akustik und Perzeption gesprochener Sprache und sie fügen den Teilbereich Transkription von Sprache hinzu.

- In der *Sprachproduktion* steuern kognitive und neuronale Prozesse Muskelbewegungen im Artikulationstrakt, z. B. Atmung, Zungen- und Lippenbewegungen.
- Die *Akustik* umfasst die physikalischen Grundlagen des Sprachschalls und seiner Übertragung.
- Bei der *Perzeption* wird der Sprachschall beim Hörer wieder in neuronale und kognitive Prozesse umgewandelt und der im Sprachsignal enthaltene Gehalt extrahiert.
- Die *Transkription* befasst sich mit der komplexen Beziehung zwischen wahrnehmbaren Sprachlauten und den zu ihrer Beschreibung verwendeten Symbolen, z. B. dem phonetischen Alphabet der IPA (*International Phonetic Association*) (IPA 1999).

Die moderne phonetische Grundlagenforschung ist häufig korpusbasiert. Harrington (2010) schreibt gleich zu Beginn seiner Einführung:

> A speech corpus is a collection of one or more digitized utterances usually containing acoustical data and often marked for annotations. The task in this book is to discuss some ways that a corpus can be analyzed to test hypotheses about how speech sounds are communicated. But why is a corpus needed for this at all? (Harrington 2010: 1)

Er beantwortet diese Frage so, dass Intuition, Introspektion und Transkription die notwendigen Voraussetzungen für phonetische Hypothesen und darauf aufbauend neue Erkenntnis seien, dass aber nur das gemessene Sprachsignal eine objektive Basis für die empirische Überprüfung dieser Hypothesen darstelle.

Im Folgenden stellt er fest, dass es auch weiterhin notwendig sei, eigene phonetische Korpora zu erstellen:

> Unfortunately, most kinds of phonetic analysis still require building a speech corpus that is designed to address a specific research question. In fact, existing large-scale corpora ... are very rarely used in basic phonetic research, partly because, no matter how extensive they are, a researcher inevitably finds that one or more aspects of the speech corpus ... are insufficiently covered for the research question to be completed. (Harrington 2010: 6)

Die erwähnten großen Korpora sind Resultat unterschiedlicher Entwicklungen:

- In der phonetischen Grundlagenforschung werden neben dem akustischen Signal auch artikulatorische Messwerte erfasst, was sehr datenintensiv ist. Dazu werden Mess- und bildgebende Verfahren aus der Medizin eingesetzt, die auf nicht-invasive Weise Vorgänge im Inneren des Körpers sichtbar machen.

- In der Sprachtechnologie-Entwicklung, insbesondere der Spracherkennung und -synthese, haben sich statistische Verfahren durchgesetzt. Diese müssen trainiert werden, wozu große und den Anwendungsbereich möglichst vollständig abdeckende Sprachdatensammlungen notwendig sind.
- In der Sprachdokumentation werden bedrohte Sprachen – oftmals einhergehend mit ethnologischer Dokumentation – phonetisch und linguistisch systematisch erfasst, dokumentiert und der Forschung zugänglich gemacht.

Diese Entwicklungen führten zu einem enormen Zuwachs an Daten, der neue Formen der Datenorganisation, -speicherung und -verfügbarkeit notwendig machte. Die zeitgleiche Entwicklung des World Wide Web zu einem weltumspannenden Kommunikationsnetz hat diese Notwendigkeit noch verstärkt.

Darüberhinaus besteht eine starke Wechselwirkung zwischen Phonetik und Sprachtechnologie: die Sprachtechnologie liefert technische Verfahren zur Speicherung und Analyse gesprochener Sprache und ermöglicht damit überhaupt erst die Verarbeitung auch großer Mengen an Sprachdaten, gerade in der Grundlagenforschung. Im Gegenzug sind Erkenntnisse der Phonetik die Basis vieler sprachtechnologischer Verfahren bzw. tragen zu deren Verbesserung bei.[1]

Weitere Aspekte sind die Forderung vieler Förderinstitutionen nach langfristigem Datenmanagement, damit aufwendig erstellte Datensammlungen auch nach Abschluss der Projektförderung verfügbar bleiben,[2] sowie die Frage nach dem kollegialen Datenaustausch und der Reproduzierbarkeit von Studien. Gut dokumentierte und nach dem Stand der Technik erstellte Korpora werden häufig ganz oder in wesentlichen Teilen mehrfach und auch zur Untersuchung verschiedener Fragestellungen genutzt. Die Bereitstellung von Rohdaten, Skripten und Auswertungsroutinen erlaubt es, Studien zu replizieren oder durch das Hinzufügen neuer Annotationen oder Daten den Nutzen oder Wert eines Korpus zu erhöhen. Die Sprachdatenbank TIMIT (Garofolo et al. 1986) ist dafür ein gutes Beispiel: ursprünglich zur Entwicklung von Spracherkennungssystemen erstellt, wurde das Datenbankdesign auf viele Sprachen und technische Kommunikationsmittel und sogar auf artikulatorische Datenbanken übertragen, die ursprünglichen Transkriptionen wurden vielfach korrigiert sowie um zusätzliche Annotationen ergänzt.

1 So kann z. B. die Wortfehlerrate bei der Spracherkennung durch phonetisches Wissen über die Artikulationsvorgänge um relativ 10–20 % verbessert werden (Richardson, Bilmes & Diorio 2003).

2 Siehe dazu die Handreichungen der DFG mit dem Titel „Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora“.

Dieser Beitrag gliedert sich in zwei Teile: im ersten Teil werden die Datenarten phonetischer und sprachtechnologischer Datenbanken beschrieben. Im zweiten Teil wird der in der Praxis relevante Prozess der Erstellung und Nutzung einer phonetischen Datenbank anhand eines konkreten Projekts präsentiert; der Fokus hierbei wird auf den fachlichen Aspekten liegen: Forschungsfrage, Datensammlung, -aufbereitung und -auswahl sowie Analyse und Interpretation.

2 Sprachdatenbanken

In der linguistischen und phonetischen Literatur wird häufig die Bezeichnung *Korpus* verwendet. Damit werden ganz allgemein entweder natürlich vorgefundene oder explizit zusammengestellte Sammlungen sprachlicher Daten bezeichnet (siehe z. B. Gippert, Himmelmann & Mosel 2006; Lemnitzer & Zinsmeister 2006). Um die enge Beziehung zwischen Phonetik und Sprachtechnologie auch terminologisch deutlich werden zu lassen, bevorzugen wir den Begriff *Sprachdatenbank* (engl. *speech database*) und verwenden ihn für eine wohlstrukturierte, auf Dauer angelegte Sammlung von digitalen Daten gesprochener Sprache, dazugehörigen Annotationen und Metadaten.

Sprachdatenbanken bestehen aus *Primär-*, *Sekundär-* und *Metadaten*. Primärdaten sind die bei der Erfassung von gesprochener Sprache erhobenen Rohdaten, Sekundärdaten davon automatisch oder manuell abgeleitete oder erstellte Mess- und Annotationsdaten. Metadaten umfassen die sonstigen erhobenen Daten bzw. beschreiben Aufbau und Struktur der Datenbank.

Primärdaten sind, abgesehen von technischer Konvertierung, prinzipiell unveränderlich. Sekundärdaten können wiederholt und mit verschiedenen Verfahren berechnet oder durch Annotation der Primärdaten erstellt werden, sie können korrigiert und um neue Daten erweitert werden. Metadaten erlauben die Beschreibung und Katalogisierung von sowie die Suche nach Datenbeständen.

Die Annotation von Primärdaten ist ein Kategorisierungsprozess: einem Signalabschnitt werden kategoriale Einheiten eines Symbolinventars zugeordnet. In der Phonetik wird dieser Vorgang als *Transkription* bezeichnet (IPA 1999: 3). Dieser Prozess ist stets mit einer Unsicherheit behaftet und somit niemals *korrekt*, sondern höchstens *plausibel*.

Erst mit der *Digitalisierung* sind das systematische Zusammenführen und die informationserhaltende Konvertierung der unterschiedlichen Datenarten technisch möglich geworden. In digitaler Form können unterschiedliche Transkriptionen miteinander verknüpft, Transkriptionen mit Zeitsignalen aligniert

und Zeitsignale miteinander synchronisiert werden. Dies eröffnet der Phonetik, der Sprachtechnologie und anderen sprachverarbeitenden Gebieten ganz neue Möglichkeiten des Zugriffs und des Erkenntnisgewinns.

2.1 Datenarten

Grundvoraussetzung phonetischer Datenbanken sind die eigentlichen Daten in digitaler Form. Jede Datenart, jedes Messverfahren hat je eigene Datenmodelle und -formate. Beschreibungen dieser Datenmodelle und -formate müssen öffentlich verfügbar sein, damit sie auch unabhängig von den Tools, mit denen sie erstellt wurden, langfristig zugänglich sind.

Zur Beschreibung der in phonetischen Datenbanken gespeicherten Daten eignet sich eine erste Unterteilung in *zeitunabhängige* und *zeitbezogene* Daten. Zeitunabhängig sind Daten, die außer der Tatsache, dass sie eine gesprochene Äußerung wiedergeben, keine Zeitinformation enthalten, z. B. der orthographische Wortlaut einer Äußerung oder eine phonetische Transkription.

Zeitbezogene Daten sind entweder Ereignis- oder Intervalldaten. Ein Ereignis hat nur einen Zeitpunkt und keine Dauer, ein Intervall hat einen Anfangszeitpunkt und eine Dauer. Beispiel für ein Ereignis ist ein Wendepunkt in einer Intonationskurve, Beispiel für ein Intervall der einem Wort, Phonem oder

Abb. 8.1: Emu WebApp Labeller mit drei Signaldarstellungen (Oszillogramm, Sonagramm, EMA-Sensormessdaten), drei Annotationsebenen (Segment, TT [*tongue tip*] und TB [*tongue back*]) sowie einer Visualisierung der Bewegungen der EMA-Sensoren in einem eigenen Fenster unten rechts.

Allophon zugeordnete Signalabschnitt einer Aufnahme (siehe Abb. 8.1). Neben dem Zeitbezug unterscheidet man die Daten nach den jeweiligen Medien. In Sprachdatenbanken sind dies in der Regel Text-, Audio- und Video- sowie Sensordaten.

Diese Daten werden soweit möglich getrennt erfasst, um sie sowohl isoliert betrachten als auch miteinander in Beziehung setzen zu können. Die eigentliche Aufgabe von Sprachdatenbanken ist, die Daten so miteinander zu verknüpfen, dass sie die Beantwortung phonetischer und sprachtechnologischer Fragestellungen erlauben. Diese an die Anforderungen der Sprachforschung und -verarbeitung angepasste Datenmodellierung unterscheidet phonetische Datenbanken von Datenbanken für andere Anwendungsgebiete oder Universaldatenbanken.

2.2 Datenmodelle für phonetische Daten

Bird & Liberman (2001) entwickelten *Annotation Graphs* als allgemeines formales Modell für Annotationen gesprochener Sprache. Annotation Graphs sind im Wesentlichen Sammlungen gerichteter azyklischer binärer Graphen. Die Knoten tragen einen eindeutigen Bezeichner und einen optionalen Zeitstempel, mit dem sie einen Zeitpunkt in einer Signaldatei angeben. Die Kanten haben einen Annotationstyp und ein Annotationslabel. Pfade innerhalb eines Annotationsgraphen sind die transitive Hülle von Kanten eines Annotationstyps; die Knoten eines Pfades sind nach Zeitstempel geordnet, eine Kante darf nicht zu einem Knoten mit einem früheren Zeitstempel führen.

Die Autoren zeigen, dass sich die in der Literatur beschriebenen oder in Annotationseditoren implementierten Annotationsnotationen für Aufnahmen gesprochener Sprache mit Annotation Graphs darstellen lassen.

Annotation Graphs eignen sich aber nur eingeschränkt für phonetische Datenbanken. Auf theoretischer Ebene ist die Beschränkung auf gerichtete azyklische Graphen in der Praxis zu eng, denn es gibt Phänomene in gesprochener Sprache, die als diskontinuierliche Elemente oder in Form zyklischer Strukturen beschrieben werden. In der Praxis ist das Fehlen eines Datenbankschemas ein Problem, da deshalb eine automatische Integritätskontrolle der Datenbank nicht möglich ist.

Ein erweitertes Datenmodell, das sowohl Annotationsebenen als auch ein explizites Datenbankschema vorsieht, ist im Emu-System realisiert (Harrington et al. 1993; Cassidy & Harrington 1996, 2001). In Emu enthält eine Annotationsebene nur Daten eines bestimmten Annotationstyps – phonetische Segmentation, phonemische Transkription, Silbe, orthographischer Wortlaut usw. Die

Annotationsebenen sind hierarchisch angeordnet, wobei es innerhalb einer Datenbank mehrere Hierarchien geben kann. Innerhalb einer Annotationsebene sind die Elemente sequenziell angeordnet, zwischen den Elementen benachbarter Ebenen gibt es Dominanzbeziehungen. Ein Schema beschreibt die in einer Sprachdatenbank vorgesehenen Annotationsebenen und in welcher quantitativen Beziehung Elemente einer Ebene zu denen benachbarter Ebenen stehen.

Das Sprachdatenbanksystem Emu-SDMS ist die von Winkelmann, Harrington & Jänsch (2017) vollständig in R implementierte neue Fassung von Emu. Emu-SDMS besteht aus der Emu WebApp zur graphischen Darstellung von Sprachdaten in einem Webbrowser sowie der eigentlichen EmuR-Sprachdatenbank und der Signalverarbeitung *wrassp*, die beide auf das lokale Dateiverzeichnis mit Audio- und Annotationsdateien zugreifen (siehe Abb. 8.1).

EmuR unterscheidet drei Klassen von Annotationselementen:

- *Item*: Element ohne Zeitbezug
- *Event*: Ereignis-Element mit einem Zeitpunkt
- *Segment*: Intervall-Element mit Anfangszeitpunkt und Dauer

Innerhalb einer Ebene sind nur Elemente eines Typs erlaubt und sie sind sequenziell angeordnet. Der Inhalt eines Annotationselements ist in seinen Labels gespeichert, jedes Annotationselement ist durch einen eindeutigen, vom System vergebenen Bezeichner identifiziert.

Das Schema legt fest, welche Ebenen miteinander verknüpft sind. Dabei sind 1 : 1-, 1 : *n*- und auch *n* : *m*-Dominanzbeziehungen möglich, außerdem kann eine Ebene in mehr als einer Hierarchie vorkommen (siehe Abb. 8.2).

Die Abfragesprache EQL (*Emu Query Language*) erlaubt die kompakte Formulierung von Abfragen von Dominanz- und Sequenzbeziehungen. Eine Abfrage in der Datenbank AE nach Silben, die ein Segment [n] auf der Ebene *Phonetic* enthalten, wird wie folgt formuliert:

```
[Phonetic == n ^ #Syllable = ~.*]
```

Der Operator ^ ist der Dominanzoperator, d. h. die Ebene *Phonetic* wird von der Ebene *Syllable* dominiert. # zeigt an, dass die Zeitinformation für die Ebene *Syllable* angezeigt werden soll – in diesem Fall bedeutet dies, dass die Zeitinformation aus der Ebene *Phonetic* in die Ebene *Syllable* propagiert wird, d. h. Anfang und Dauer der ganzen Silbe zurückgegeben werden sollen; für die Silben gelten keine weiteren Einschränkungen.

Der Dominanzoperator ^ ist rekursiv, d. h. er kann über mehrere Ebenen einer Hierarchie berechnet werden.

Abb. 8.2: Schema der Datenbank AE (*Australian English*) mit mehreren Hierarchien und den beiden zeitbasierten Annotationsebenen *Phonetic* der Klasse *Segment* und *Tone* der Klasse *Event* aus Harrington (2010: 99).

Abfragen in der Emu-Datenbank ergeben Segmentlisten der Form

```
AUDIOFILE ELEMENT BEGIN DURATION
```

wobei `AUDIOFILE` die Audiodatei der Äußerung ist, `ELEMENT` das Annotationselement der gewählten Ebene, und `BEGIN` und `DURATION` die Zeitangaben der zeitbezogenen Annotationsebene der aktuell ausgewählten Hierarchie.

Die direkte Einbindung in die Statistiksoftware R erlaubt in Kombination mit dem Signalverarbeitungspaket *wrassp* statistische Auswertungen des Datenbestands sowie die Aufbereitung der Daten für die Visualisierung.

Neben Emu gibt es weitere Ansätze für phonetische Datenbanken. So beschreiben Draxler & Kleiner (2015) eine phonetische Datenbank auf der Basis eines relationalen Datenbanksystems. Damit sind auf einfache Weise Abfragen auch über mehrere Sprachdatenbanken möglich und auch nichtsprachliche Daten wie z. B. Orts- und Signaldateiangaben können redundanzfrei gespeichert werden. Allerdings müssen die Abfragen in SQL formuliert werden und

Signalverarbeitung, statistische Auswertung und Visualisierung müssen außerhalb des Datenbanksystems erfolgen.

2.3 Textdaten

Textdaten sind Lesetexte, Annotationen, orthographische, breite phonemische und enge phonetische Transkriptionen, Datentabellen und statistische Rohdaten, frei formatiert oder mit definierter Struktur. Sie sind vorgegeben bzw. wurden durch manuelle oder automatische Annotation erzeugt. Technisch sollten Textdaten in Unicode und UTF-8-Kodierung vorliegen, die Struktur von Textdokumenten sollte öffentlich definiert oder selbstbeschreibend sein.[3,4]

Die IPA empfiehlt Transkriptionen auf mindestens zwei Ebenen: breite phonemische Transkription der Wörter in Zitierform und eine enge phonetische Transkription (IPA 1989: 81), dazu kommt üblicherweise noch eine orthographische Transkription (Gibbon, Moore & Winski 1997: 152). Je nach Transkriptionstyp werden verschiedene, an die jeweilige Aufgabe angepasste Editoren verwendet. Die orthographische Transkription soll möglichst rasch erstellt werden, phonetische Fachkenntnisse sind in der Regel nicht notwendig. Die Organisation der Transkriptionsarbeit via Crowdsourcing und das Verwenden von webbasierten Editoren mit entweder einem Standard-Audioplayer oder einem einfachen Oszillogramm haben sich bewährt.

Abbildung 8.3 zeigt die 2D-Ansicht für lange Signaldateien sowie das Editierfenster des webbasierten Editors OCTRA (Pömp & Draxler 2017).

Editoren für die phonetische Mehr-Ebenen-Annotation bieten in der Regel eine partiturartige graphische Darstellung. Bekannte Editoren sind Praat (Boersma & Weenink 1996), EXMARaLDA (Schmidt & Wörner 2005), ELAN[5] (Sloetjes, Russel & Klassmann 2007) und Emu Webapp (Winkelmann, Harrington & Jänsch 2017). Abbildung 8.1 zeigt als Beispiel den Emu WebApp Labeller.

Diese Editoren unterscheiden sich z. T. erheblich in ihrer Bedienung und ihrem Funktionsumfang. Die damit erzeugten Annotationsdaten sind nur teilweise kompatibel, so dass beim Datentransfer Informationsverlust auftreten kann. Für eine Diskussion dieses Themas siehe z. B. Schmidt et al. (2009) oder Draxler et al. (2011).

---

3 Das IPA-Alphabet zur Wiedergabe phonetischer Zeichen ist integraler Bestandteil der Unicode-Zeichentabelle und kann somit in Unicode-kompatibler Software verwendet werden.

4 Die *Text Encoding Initiative* (TEI) hat einen Standard zur Transkription gesprochener Sprache vorgeschlagen, siehe http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html (letzter Zugriff: 7. 11. 2017).

5 ELAN ist primär für die Annotation von Videos gedacht.

Abb. 8.3: OCTRA Annotationseditor mit einem 2D-Oszillogramm zur Darstellung langer Audiodateien (links) und mit überlagertem Transkriptionseditor für Segmente (rechts).

2.4 Audiodaten

Digitale Audiodaten erfassen das akustische Signal. Die Abtastrate (engl. *sampling rate*) gibt an, wie viele diskrete Werte pro Zeiteinheit erfasst werden, die Quantisierung, wie viele Werte unterschieden werden können. Üblicherweise verwendet man für Sprachdaten Abtastraten von mindestens 16.000 Messpunkten pro Sekunde (abgekürzt 16 kHz) und eine Quantisierung von mindestens 16 Bit. Sprachaufnahmen erfolgen über ein oder mehrere Mikrofone, die über einen Analog/Digital-Konverter an ein Aufnahmegerät oder einen Rechner angeschlossen sind. Ziel von Sprachaufnahmen ist in der Regel ein unter den gegebenen Umständen und für die geplante Untersuchung optimales Sprachsignal zu bekommen. Dazu stehen verschiedene Mikrofontypen zur Verfügung, die Aufnahmen können in kontrollierten Bedingungen im Studio oder im Feld erfolgen.[6]

Für standardisierte Sprachaufnahmen, in denen z. B. vorbereitete Sätze vorgelesen werden, ist der Einsatz von spezieller Software sinnvoll. Die Software SpeechRecorder (Draxler & Jänsch 2004) führt skriptgesteuert Sprachaufnahmen durch und schreibt jede Aufnahme automatisch in eine separate Datei. Damit sind teil- oder vollautomatisch ablaufende Aufnahmesitzungen möglich, ein nachträgliches Schneiden von Signalen kann weitgehend vermieden werden (siehe Abb. 8.4).

Audiodaten sollten entweder gar nicht oder verlustfrei komprimiert werden. Verlustbehaftete Kompression entfernt aus dem Sprachsignal für den Menschen nicht wahrnehmbare Anteile, diese können aber in automatischen Verfahren durchaus relevant sein.

6 Eine Einführung in Aufnahmetechnik und -situationen gibt Draxler (2008: 132–169).

Abb. 8.4: SpeechRecorder Sprecheransicht (links) und Aufnahmeleiteransicht (rechts). In der Aufnahmeleiteransicht sind zusätzlich ein Oszillogramm zur Beurteilung des Mikrofonpegels sowie die Liste aller schon aufgenommenen bzw. noch ausstehenden Aufnahme-Items zu sehen.

2.5 Videodaten

Videoaufnahmen erfassen sichtbare Aspekte gesprochener Sprache. Das reicht von Gesamtaufnahmen einer Aufnahmesituation über Halbtotale und Gesichtsaufnahmen bis hin zu Nahaufnahmen der Lippen. Üblich sind aktuell Videoaufnahmen mit einer Bildgröße von mindestens HD-Auflösung (d. h. 1920 × 1080 Pixel) mit einer Bildwiederholrate von 30 Bildern pro Sekunde; für spezielle Anwendungen sind auch Bildwiederholraten von weit über 100 Bildern pro Sekunde notwendig.

Digitales Video wird in der Regel verlustbehaftet komprimiert, um die Datenrate und den damit verbundenen Speicherbedarf zu begrenzen.

2.6 Sensordaten

Sensordaten werden vor allem im Bereich der Sprachproduktion erhoben. Damit werden Strukturen und Bewegungen im Körperinneren erfasst und teilweise auch direkt sichtbar gemacht. Aus den komplexen Muskel- und Artikulatorbewegungen sind Rückschlüsse auf die neuronale Ansteuerung und muskuläre Koordination sowie auf Art und Organisation des mentalen Lexikons möglich.

Die im Folgenden beschriebenen Sensordaten sind von besonderer Relevanz für moderne Sprachdatenbanken.

2.6.1 Elektropalatographie

Bei der Elektropalatographie erfassen in einer Matrix angeordnete Elektroden in einem künstlichen Gaumen den Kontakt mit der Zunge. Damit lassen sich

Abb. 8.5: Künstlicher Gaumen mit 62 Elektroden für die Elektropalatographie [links, aus Gibbon & Crampin (2001: 98)] und schematisches Elektropalatogramm des /s/ in der Äußerung /a s a/ (rechts).

insbesondere Konsonanten gut erfassen, weil diese durch Enge- oder Verschlussbildung im Artikulationstrakt gebildet werden.

Die Abtastrate beträgt bis zu 200 Hz, die Quantisierung mindestens 64 Bit. Mit dem künstlichen Gaumen kann nur der Bereich des harten Gaumens abgedeckt werden. Der Zungenkontakt ganz vorne im Artikulationstrakt, an Lippen oder Zähnen, oder ganz hinten im Bereich des Velums und des Rachens kann nicht erfasst werden (siehe Abb. 8.5).

2.6.2 Laryngographie und Laryngoskopie

Bei der Laryngographie werden die Schwingungen der Stimmlippen im Kehlkopf mit Elektroden auf der Haut gemessen (siehe Abb. 8.6), bei der Laryngoskopie werden sie gefilmt. Dazu wird eine kleine Videokamera mit Lichtquelle vom Rachenraum aus auf die Stimmlippen gerichtet. Erwachsene Männer haben eine Grundfrequenz von 50 bis 150 Hz, Frauen von 200 bis 300 Hz, Kinder deutlich darüber. Daher sind sehr hohe Bildwiederholraten von 100 bis über 2.000 Hz notwendig.

Häufig werden in Laryngographie-Videos die Konturen der Stimmlippen automatisch ermittelt oder manuell erfasst und in Form von Vektordaten gespeichert. Damit ist eine erhebliche Reduktion des Datenumfangs und eine präzise Messung der Geschwindigkeit und Beschleunigung einzelner Punkte auf den Stimmlippen möglich.

Abb. 8.6: Oszillogramm und Laryngographensignal der Äußerung /p a t a x/.
Die stimmlosen Plosive /p/ und /t/ und der stimmlose Frikativ /x/ sind im Oszillogramm deutlich zu erkennen, im Laryngogramm nicht. Die beiden /a/-Vokale sind, da sie stimmhaft sind, als synchrone Schwingungen sowohl im Oszillogramm als auch im Laryngogramm zu sehen (aus Draxler 2008: 71).

2.6.3 Echtzeit Magnet-Resonanz-Tomographie

Magnet-Resonanz-Tomographie (MRT, engl. *magnet resonance imaging MRI*) ist ein bildgebendes Verfahren zur Darstellung des weichen Gewebes im Körper. Dabei wird der Körper schrittweise in einzelnen Schichten erfasst, aus denen zweidimensionale Schnittbilder oder auch dreidimensionale Darstellungen berechnet werden können.

In der Phonetik wird MRT zur Darstellung des Artikulationstrakts beim Sprechen verwendet. Damit lassen sich die Positionen der Zunge, des Velums, der Lippen und des Kiefers visualisieren. In vielen MRT-Geräten muss der Sprecher liegen. Das beeinflusst die Geometrie des Artikulationstrakts.

Werden mehrere MRT-Bilder in so kurzer Zeit nacheinander aufgenommen, dass sie als flüssig ablaufendes Video betrachtet werden können, spricht man von Echtzeit-MRT. Aktuell sind Bildwiederholraten von über 50 Bildern pro Sekunde möglich. In der Regel sind in MRT-Filmen Details mit einer Kantenlänge von knapp 2 mm zu erkennen.

MRT-Aufnahmen sind sehr aufwendig und teuer, sie können nur an wenigen Labors weltweit durchgeführt werden. Da die Geräte sehr laut sind und ein starkes Magnetfeld erzeugen, sind synchrone Sprachaufnahmen nur mit nichtmetallischen Mikrofonen und starken Störgeräuschen möglich.

Abb. 8.7: MRT-Video-Standbild mit überlagerter berechneter Kontur des Artikulationstrakts sowie Oszillogramm und Sonagramm (Narayanan et al. 2014: 1310).

Wie bei anderen bildgebenden Verfahren kann durch Tracken der Konturen in den einzelnen Videobildern eine datenreduzierte Vektordarstellung erzeugt werden, die kinetische Daten einzelner Punkte bei der Artikulation liefert (siehe Abb. 8.7).

2.6.4 Ultraschall

Mit Ultraschall oder *Sonographie* kann man Schallreflexionen an den Übergängen zwischen Gewebe und Luft messen. In der Phonetik wird Ultraschall zur nicht-invasiven Erfassung der Zungenbewegungen verwendet; dazu wird die Sonde auf der Haut an der Unterseite des Unterkiefers angesetzt. Ultraschallaufnahmen gelten als ungefährlich. Das korrekte Positionieren der Sonde ist für ein klares Bild wichtig. Die Zunge kann bis fast zur Zungenspitze erfasst werden (siehe Abb. 8.8). Die Abtastrate kann bei reduzierter räumlicher Auflösung über 100 Bilder pro Sekunde betragen, wird aber häufig auf die Bildrate beim Videoexport, d. h. 25–30 Bilder pro Sekunde, beschränkt. Der geringe technische Aufwand, die gute Softwareunterstützung sowie die einfache Handhabung haben zu einer raschen Verbreitung von Ultraschall geführt. Synchrone Audioaufnahmen sind möglich, tragbare Ultraschallgeräte erlauben Aufnahmen im Feld.

Abb. 8.8: Ultraschallaufnahme der Zungenbewegung. Die Zungenkontur ist durch die helle Linie deutlich sichtbar. Unten links ist die Sensorposition eingeblendet, unten rechts das Oszillogramm der Äußerung (aus Stone 2005: 25).

2.6.5 Elektromagnetische Artikulation

Bei der Bewegung von Spulen in einem Magnetfeld wird in den Spulen ein Strom induziert. Bei der elektromagnetischen Artikulographie (EMA) befinden sich kleine Spulen auf Zunge und Lippen des Sprechers. Moderne EMA-Geräte erlauben eine freie Bewegung des Kopfes im magnetischen Feld und eine Erfas-

Abb. 8.9: EMA-Zungenkonturen in der Äußerung „tote“. Die obere Kurve ist der harte Gaumen, die Mundöffnung ist links. Die untere viergliedrige Linie ist die Zungenkontur bei gehaltenem /o:/, die mittlere Linie die Zungenkontur für die Phoneme /t/ (links) und /o/ (rechts).

sung der Bewegung der Spulen in fünf Dimensionen (*x*,*y*,*z*-Koordinaten sowie Rotation in zwei Ebenen). Die EMA liefert Positionsdaten der einzelnen Spulen, aus denen zweidimensionale (siehe Abb. 8.9) oder räumliche Darstellungen der Artikulation berechnet werden. Mit der EMA können Details im Bereich eines Millimeters gemessen werden, die Abtastrate beträgt bis zu 1.250 Hz. EMA-Aufnahmen sind aufwendig und erfordern geschultes Personal.

2.6.6 Eyetracking

Beim Eyetracking werden Blickrichtung, -bewegung und Veränderung der Pupille gemessen. Die Alignierung der Bewegung mit dem akustischen Sprachsignal erlaubt Rückschlüsse auf mentale Prozesse bei der Sprachverarbeitung, sowohl bei der Produktion als auch der Perzeption. Huettig, Rommers & Meyer (2011) geben einen Überblick der Studien mit Eyetracking, Holmqvist, Nyström & Mulvey (2012) diskutieren die Zuverlässigkeit von Eyetracking-Messdaten.

Die Sensoren für das Eyetracking sind entweder in der Nähe des zu erfassenden Bildschirms angeordnet, oder sie befinden sich in einer speziellen Brille. Typische Abtastraten liegen im Bereich von 30 bis 300 Hz, der anvisierte Punkt wird mit *x*,*y*-Koordinaten angegeben, woraus sich bei bekannter Entfernung des gemessenen Auges von der betrachteten Bildfläche der Winkel der Bewegung berechnen lässt. Wichtige Angaben zur Qualität der Messung sind die Abweichung (engl. *accuracy* oder *offset*), die angibt, wie weit der anvisierte Punkt vom Ziel abweicht, und die Streuung (engl. *precision*), die angibt, um welchen Betrag die Messwerte bei unveränderter Blickrichtung streuen; beide werden in Winkelgraden angegeben.

2.6.7 Röntgenbilder und -filme

Wegen ihrer Gesundheitsgefährdung werden Röntgenaufnahmen bzw. das damit verwandte Verfahren X-Ray Microbeam (XRMB) nur noch in speziellen Situationen, z. B. zur Vorbereitung und Nachsorge von Operationen, durchgeführt. Es gibt allerdings viele historische Röntgenaufnahmen, sowohl Stand- als auch Bewegtbilder, die digitalisiert wurden und nun zu Lehr- und Forschungszwecken genutzt werden (Abb. 8.10).[7]

---

7 Diese Abbildung sowie die EMA-Darstellungen in Abbildung 8.9 wurden freundlicherweise von Phil Hoole vom Institut für Phonetik und Sprachverarbeitung der LMU München zur Verfügung gestellt.

Abb. 8.10: Röntgenaufnahme des Mundraums eines Sprechers bei der Artikulation des Diphthongs /a/ und Sonagramm der Äußerung *It's ten below outside*.

XRMB ist ein Verfahren, bei dem ein extrem schmaler Röntgenstrahl verwendet wird, um 2–3 mm große Goldkügelchen, die in Längsrichtung mittig auf die Zunge, die Lippen, den Kiefer und den weichen Gaumen geklebt werden, in ihrer Bewegung zu erfassen. Durch den sehr schmalen Röntgenstrahl kann die Belastung der aufgenommenen Person stark reduziert werden. Die Kügelchen sind im Röntgenbild deutlich erkennbar, so dass durch Interpolation zwischen den Messpunkten die Kontur der Zunge berechnet werden kann.[8]

8 XRMB-Aufnahmen werden seit 1993 praktisch nicht mehr durchgeführt.

2.7 Beispiele phonetischer Datenbanken

Seit der Veröffentlichung von TIMIT wurden laufend weitere Sprachdatenbanken für die phonetische Grundlagenforschung und die Entwicklung von Sprachtechnologie erstellt. Um die Sicht- und Verfügbarkeit dieser Sprachdatenbanken zu verbessern, wurden Zentren wie das *Linguistic Data Consortium* (LDC) in den USA, die *European Language Resources Association* (ELRA) in Europa, das *Bayerische Archiv für Sprachsignale* (BAS) und das *Hamburger Zentrum für Sprachkorpora* (HZSK) in Deutschland und viele weitere ähnliche weltweit gegründet. In sog. Repositories katalogisieren sie die Sprachdatenbanken, bieten Such- und Blätterfunktionen im Datenbestand an und erlauben das Herunterladen bzw. Lizenzieren von Sprachdatenbanken. Viele Zentren bieten darüberhinaus web-basierte sprachtechnologische Dienste an. Ein Beispiel dafür ist die automatische phonetische Segmentation von Sprachaufnahmen auf den Webseiten des BAS (http://clarin.phonetik.uni-muenchen.de/BASWebServices/ [letzter Zugriff: 7. 11. 2017]).

Die weitaus meisten der in Repositories verfügbaren Datenbanken sind akustische Sprachdatenbanken. Abbildung 8.11 zeigt beispielhaft die Webseite des BAS Repositories. Dort sind mehr als 40 Sprachdatenbanken aufgelistet, die größtenteils am Institut für Phonetik und Sprachverarbeitung der LMU München erstellt wurden. Zunehmend kommen aber auch von Dritten erstellte Sprachdatenbanken hinzu, denn das Repository hält diese Daten auch nach Auslaufen der jeweiligen Projektfinanzierung vor.

Abb. 8.11: Webseite des BAS Repository mit Sprachdatenbanken (http://clarin.phonetik.uni-muenchen.de/BASRepository/ [letzter Zugriff: 7. 11. 2017]).

Im Gegensatz zu den akustischen Sprachdatenbanken gibt es nur sehr wenige verfügbare Sprachdatenbanken mit artikulatorischen Daten. Die im Folgenden aufgeführten vier Sprachdatenbanken sind daher in der Forschung vielgenutzte und trotz ihres teilweise schon weit zurückliegenden Entstehungsdatums immer noch aktuelle Ressourcen. So nennt z. B. Richmond, Hoole & King (2011) in einem Überblick neun Phonetik- und Sprachtechnologiefelder, in denen Arbeiten auf der Basis der 1999 erstellten MOCHA EMA-Sprachdatenbank publiziert wurden.

Diese vier Sprachdatenbanken enthalten hauptsächlich englische Sprachdaten. Vergleichbare öffentlich verfügbare Sprachdatenbanken mit primär deutschen Sprachaufnahmen sind uns nicht bekannt.

2.7.1 X-Ray Microbeam

Für die *X-Ray Microbeam Speech Production*-Sprachdatenbank (Westbury, Turner & Dembrovski 1994) wurden in hauptsächlich zwei Phasen von 30 bzw. 10 Monaten in den Jahren 1987 bis 1991 XRMB-Aufnahmen durchgeführt.[9] In die Datenbank aufgenommen wurden die Daten von insgesamt 57 Personen und je ca. 18 Minuten Sprachmaterial. Neben den Messdaten sind die gelesenen Sätze sowie ihre phonetischen Umschriften verfügbar.

Die XRMB-Sprachdatenbank wurde von Anfang an als offene Ressource für die Wissenschaft erstellt. Forscherinnen und Forscher sollten diese sehr aufwendigen Aufnahmen ohne Einschränkungen nutzen können. Auch war es explizites Ziel, vielen verschiedenen Forschergruppen Zugang zum Aufnahmegerät zu ermöglichen, um die Verbreitung und gemeinsame Nutzung der Daten zu fördern.

2.7.2 MOCHA EMA

Die MOCHA EMA-Sprachdatenbank wurde im November 1999 am Queen Margaret University College Edinburgh aufgenommen (Wrench & Hardcastle 2000). Sie enthält synchrone Audio-, Elektropalatographie-, Laryngographie- und EMA-Aufnahmen von 460 Sätzen aus der britischen TIMIT-Sprachdatenbank. Diese

---

9 Das Kapitel „A Short History of the UW XRMB Facility“ im Bericht von Westbury, Turner & Dembrovski (1994) zeigt eindrucksvoll, dass schon Entwurf und Realisierung des XRMB-Geräts mehrjährige Projekte waren, mit denen technologisches Neuland betreten wurde.

Datenbank ist auf den Seiten der Universität Edinburgh unter http://data.cstr.ed.ac.uk/mocha/ (letzter Zugriff: 7. 11. 2017) frei verfügbar.

Von je einer Sprecherin und einem Sprecher gibt es alle 460 gelesenen Sätze mit allen Messdaten; diese wurden auch manuell überprüft und wo notwendig korrigiert. Von 10 weiteren Personen gibt es ausreichend viele Daten, so dass sie ebenfalls in die Datenbank aufgenommen wurden. Geplant war eine weitaus größere Anzahl an Aufnahmen. Diese wurde aber nicht erreicht, weil nur wenige Personen bereit waren, die unangenehme Prozedur des Anpassens eines künstlichen Gaumens zu durchlaufen, und weil die Organisation der Aufnahmen zeitlich deutlich aufwendiger war als ursprünglich geplant.

2.7.3 mngu0

Die mngu0-Sprachdatenbank ist eine Ergänzung und Verbesserung der MOCHA EMA-Sprachdatenbank. Ausgangspunkt ist die Feststellung in Richmond, Hoole & King (2011), dass a) für empirische Analysen zusätzliche Daten benötigt werden, b) die Vergleichbarkeit von Daten verschiedener Quellen nur bedingt möglich ist, und dass c) bekanntgewordene technische Fehler korrigiert werden und neue Messverfahren zum Einsatz kommen können.

Die mngu0-Sprachdatenbank wird EMA-Sensormessdaten mit synchronen Audioaufnahmen und Videoaufnahmen der unteren Gesichtshälfte, volumetrische MRT-Scans des Lautinventars der Sprecher sowie 3D Modellierungen der Zähne im Unter- und Oberkiefer umfassen. Die EMA-Aufnahmen wurden am Institut für Phonetik und Sprachverarbeitung der LMU München durchgeführt, und sie bestehen aus 2.000 an zwei aufeinanderfolgenden Tagen gelesenen Sätzen. Die 1.354 Aufnahmen des ersten Tages, bestehend aus aufbereiteten EMA-Spuren und den Audioaufnahmen, sind als erste Version der mngu0-Sprachdatenbank im Internet unter http://www.mngu0.org (letzter Zugriff: 7. 11. 2017) frei verfügbar.

2.7.4 USC-TIMIT

Die USC-TIMIT ist eine noch in Aufbau befindliche Sprachdatenbank mit MRT-Bewegtbildern, EMA-Daten und Audioaufnahmen (Narayanan et al. 2011, 2014). Um die Vergleichbarkeit zu den bestehenden artikulatorischen Datenbanken zu gewährleisten wurden die 460 Sätze der TIMIT-Sprachdatenbank verwendet. Laut Narayanan et al. (2014) wurden bislang je fünf Sprecherinnen und Sprecher aufgenommen und ihre Daten für die Nutzung aufbereitet. Ge-

plant sind weitere Aufnahmen mit Sprechern von anderen Muttersprachen als amerikanischem Englisch.

Daten und Auswertungssoftware sind nach einer Online-Registrierung frei verfügbar (http://sail.usc.edu/span/usc-timit/).

3 Workflow bei der Erstellung einer Sprachdatenbank

Dieser Abschnitt beschreibt die Erstellung einer Sprachdatenbank anhand des konkreten Beispiels der Sprachdatenbank *Brothers*. Diese Sprachdatenbank entstand im Rahmen der Dissertation von Hanna Feiser, in der untersucht wurde, inwieweit sich die Stimmen von Brüdern voneinander unterscheiden (Feiser 2015). Es gibt viele Untersuchungen zur Stimmähnlichkeit von Zwillingen, jedoch bislang noch keine zu Brüderpaaren. Bei Zwillingsuntersuchungen ist die genetische Ausstattung und damit verbunden die Physiognomie innerhalb eines Paares weitgehend identisch – Unterschiede in der Stimme sind also auf soziale Faktoren zurückzuführen. Bei Brüderpaaren ist die genetische Ausstattung verschieden, ebenso die Physiognomie, aber dennoch werden – das zeigt die Alltagserfahrung – die Stimmen von Familienangehörigen häufig miteinander verwechselt, gerade unter akustisch ungünstigen Bedingungen wie z. B. am Mobiltelefon. Ziel der Dissertation war daher, mögliche sprachrelevante soziale Faktoren konstant zu halten und die Beziehung zwischen akustischen Merkmalen und Sprecheridentifikation zu untersuchen. Diese Frage ist gerade in der forensischen Praxis relevant, da hier eine möglichst sichere Zuordnung einer Sprachaufnahme zu einer Person wünschenswert ist und Brüderpaare deutlich häufiger sind als Zwillingspaare.

Die Erstellung einer Sprachdatenbank gliedert sich nach Schiel et al. (2003) in die Phasen *Spezifikation*, *Vorbereitung*, *Datensammlung*, *-aufbereitung*, *Annotation*, *Dokumentation*, *Validierung* und *Distribution*. Diese Phasen sind in Abbildung 8.12 dargestellt und werden in entsprechender Abfolge in diesem Abschnitt beschrieben. Die Sprachdatenbank *Brothers* eignet sich besonders gut als Fallbeispiel, weil sie alle vier in Abschnitt 1 genannten Teilgebiete der Phonetik berührt:

- *Produktion*: Es wird sowohl gelesene als auch Spontansprache von Brüderpaaren aufgezeichnet.
- *Akustik*: Die Aufnahmen erfolgen in zwei akustischen Qualitäten im Studio bzw. über Mobiltelefon.
- *Perzeption*: Die Stimmähnlichkeit wird mit einem akustischen Perzeptionsexperiment ermittelt.

Abb. 8.12: Phasenmodell zur Erstellung von Sprachdatenbanken nach Schiel et al. (2003).

– *Transkription*: Die automatisch erstellte Segmentation bzw. die berechneten Merkmale werden manuell überprüft und ggf. angepasst.

Außerdem ist die Sprachdatenbank über ein Repository für Forschungszwecke frei verfügbar.

3.1 Spezifikation

Für die Sprachdatenbank *Brothers* soll sowohl gelesene als auch spontane Dialogsprache von mindestens zehn Brüderpaaren in zwei technischen Qualitäten aufgenommen werden. Die folgenden demographischen Faktoren sollen soweit möglich konstant gehalten werden:

– *regionale Herkunft*: Die Brüderpaare sollen aus der Region München stammen.
– *örtliche Konstanz*: Die Familien der Brüder sollen möglichst seit mindestens einer Generation im selben Wohnort leben.
– *gemeinsames Aufwachsen*: Die Brüder sollen im selben Haushalt aufgewachsen sein und immer noch in engem Kontakt stehen.
– *Alter und Altersabstand*: Die Brüder sollen mindestens 18 Jahre alt sein, der Altersabstand solle zwischen 2 und 10 Jahren betragen.

Die Sprecher werden über persönliche Kontakte sowie Teilnahmeaufrufe über soziale Medien und Mailinglisten geworben.

Die Aufnahmen sollen im Tonstudio des Instituts für Phonetik erfolgen. Zur akustischen Trennung werden die Sprecher in getrennten Räumen ohne Sichtkontakt aufgenommen. Die Aufnahmen erfolgen synchron in hoher Signalqualiät (44,1 kHz Abtastrate, 16 Bit lineare Quantisierung) mit einem Großmembranmikrofon (Neumann TLM 103 P48) sowie in Telefonqualität über Mobiltelefon (Nokia 1680C-2), wobei das Telefonsignal nach der Aufnahme auf 8 kHz Abtastrate und 16 Bit lineare Quantisierung konvertiert wird. Zusätzlich werden Formant- und F0-Tracks mit der Software ASSP berechnet.

Das Sprachmaterial besteht aus:

- 80 Minimalpaaren in Trägersätzen der Form *Anna hat … gesagt*.
- 100 zu lesenden Sätzen, den sog. *Berliner Sätzen*, die alle Phonemkombinationen mit Vokalen des Deutschen enthalten.
- einem spontanen Dialog zu Ausschnitten aus einer Folge der Fernsehkrimiserie *Tatort*. Die Brüder bekommen je unterschiedliche Ausschnitte zu sehen, damit sich ein Dialog entwickeln kann.

Die Aufnahmen erfolgen mit der Software SpeechRecorder.

Die Sprecher hören einander während des Lesens der Sätze nicht, so dass hier keine gegenseitige Beeinflussung gegeben ist.

3.2 Vorbereitung

Die Vorbereitung umfasst die Punkte Einverständniserklärung, Einrichtung der Tontechnik sowie die Vorbereitung des Aufnahmematerials.

Die Sprecher wurden vor Beginn der Aufnahmen über den Zweck der Aufnahmen sowie die Verwendung der Sprachdaten aufgeklärt, insbesondere auch darüber, dass die Aufnahmen in Form einer Sprachdatenbank auch anderen für Forschungszwecke zugänglich gemacht werden.

Das Tonstudio des Instituts für Phonetik und Sprachverarbeitung hat eine schallgedämmte Aufnahmekabine, einen reflexionsarmen Raum sowie den Kontrollraum. Für die Aufnahmen wurden die Kabine und der Kontrollraum verwendet. Im Fenster der Aufnahmekabine ist ein Monitor angebracht, auf dem die Sprecheransicht eines SpeechRecorder-Skripts sichtbar ist. Die Steuerung der SpeechRecorder-Aufnahme erfolgt von der Aufnahmeleiterin im Kontrollraum.

Für die Aufnahmen über Mobiltelefon wurde ein ISDN-Server eingerichtet, der für eine Sitzung von beiden Telefonen angerufen wurde. Für diesen Anruf wurden die beiden ISDN-Kanäle miteinander verbunden.

Das Aufnahmematerial wurde in ein SpeechRecorder-Aufnahmeskript importiert; dieses Skript umfasst die drei Abschnitte Einleitung, Minimalpaare

und Sätze in dieser Reihenfolge. Innerhalb der Abschnitte Minimalpaare und Sätze werden die einzelnen zu lesenden Prompts in zufälliger Reihenfolge präsentiert.

3.3 Datensammlung

Die Aufnahmen fanden im Zeitraum Oktober 2012 bis November 2013 statt.

Die Aufnahmesitzungen waren dreigeteilt: zuerst saß ein Bruder in der Aufnahmekabine und hat die Sätze gelesen, während der zweite sich im Kontrollraum verschiedene Ausschnitte aus einer *Tatort*-Folge angeschaut hat. Danach haben beide ihren Platz getauscht. Im abschließenden Dialog haben sie sich maximal zehn Minuten über die gesehenen Filmausschnitte unterhalten. Während der Aufnahmen hatten sie keinen Sichtkontakt.

Die Sprachaufnahmen der gelesenen Sprache erfolgten in der Kabine des Tonstudios. Die Aufnahmeleiterin hat den Fortgang der Aufnahmen über den Aufnahme-PC gesteuert. Der Sprecher hat die zu sprechenden Sätze von einem Monitor im Fenster der Kabine abgelesen. Versprecher usw. wurden sofort korrigiert, indem die Aufnahme wiederholt wurde.

Die SpeechRecorder-Aufnahmesoftware hat die mit dem Studiomikrofon aufgezeichneten gelesenen Sätze in je eigene Dateien geschrieben. Der Dialog wurde mit der Software Audacity aufgezeichnet. Auf dem ISDN-Server wurde die gesamte Sitzung als eine lange Audiodatei gespeichert.

3.4 Datenaufbereitung

Die Sprachaufnahmen auf dem ISDN-Server wurden manuell entsprechend der gelesenen Sätze geschnitten und verlustfrei in das WAV-Format mit 8 kHz Abtastrate mit 16 Bit Quantisierung konvertiert. Die mit Audacity erstellten Dialogaufnahmen wurden manuell in je zwei Aufnahmekanäle geteilt, um die jeweiligen Beiträge der beiden Sprecher getrennt voneinander bearbeiten zu können.

Für jede Audiodatei eines gelesenen Satzes wurde eine separate Textdatei mit gleichem Dateinamen und der Extension .txt angelegt, die den Wortlaut der Äußerung in normalisierter Schreibweise enthielt.

Die Dialogaufnahmen wurden nachträglich geringfügig überarbeitet, um z. B. Passagen, in denen die Aufnahmeleiterin zu hören war, zu entfernen.

Für die akustische Analyse der gelesenen Sätze wurden sowohl für die Studio- als auch die Mobiltelefonaufnahmen mittels der Emu-Signalverarbeitung die ersten vier bzw. drei Formanten sowie die Grundfrequenz f0 berech-

net. Die Formantwerte für die Telefonaufnahmen wurden anschließend in Emu manuell korrigiert.

Alle Daten wurden anschließend in das Aufnahmeverzeichnis auf dem Projektrechner im Institutsnetzwerk kopiert.

3.5 Annotation

Die Annotation erfolgte in zwei Schritten: zunächst wurden die Sätze mit dem Webdienst WebMAUS automatisch segmentiert. Ergebnis der Segmentation waren zum einen TextGrid-Dateien mit den drei Annotationsebenen ORT, KAN und MAU für die normalisierte Orthographie, die kanonische Form und die phonemische Segmentation.

Anschließend wurde die automatisch erstellte Segmentation mit der Software Praat manuell überprüft und ggf. angepasst. Die Segmentation der Aufnahmen über das Studiomikrofon erforderte nur wenige Korrekturen, die Segmentation der Telefonaufnahmen war dagegen deutlich schlechter. Der Grund dafür ist, dass die akustischen Modelle von WebMAUS nur mit Studioaufnahmen trainiert wurden. Die akustischen Eigenschaften von Mobiltelefondaten unterscheiden sich deutlich von denen von Studioaufnahmen, mit der Folge, dass die automatische Segmentation für Mobiltelefonaufnahmen sowohl für das Setzen der Labels als auch für die Segmentgrenzen schlechtere Ergebnisse liefert.

Neben der Segmentation wurden auch zwei Perzeptionsexperimente durchgeführt. Im ersten Experiment wurden phonetisch geschulte Hörer in einem ABX-Diskriminationstest befragt, im zweiten phonetisch nicht geschulte Hörer

Abb. 8.13: Trainingsphase (links) und Testphase (rechts) des zweiten Perzeptionsexperiments. Beim Training assoziiert die Teilnehmerin eine Stimme mit einer Farbe oder einer Form, beim Test muss sie die gehörte Stimme der entsprechenden Farbe oder Form zuordnen (aus Feiser 2015).

in einem Zuordnungstest; dieses zweite Experiment wurde parallel sowohl in der kontrollierten Umgebung des Tonstudios als auch in frei wählbarer Umgebung durchgeführt.

Die Sprachaufnahmen für das erste Perzeptionsexperiment stammen von Sprechern aus dem ripuarischen (mittelfränkischen) Sprachraum aus früheren Aufnahmen; diese sind ebenfalls Bestandteil der Sprachdatenbank. Für das zweite Perzeptionsexperiment wurden die am Institut für Phonetik und Sprachverarbeitung durchgeführten Aufnahmen mit bairischen Sprechern verwendet.

Das erste Perzeptionsexperiment wurde mit Praat durchgeführt, das zweite mit der Online-Experiment-Software percy (Draxler 2011). Abbildung 8.13 zeigt die Eingabemasken der Trainings- bzw. der Testphase des Online-Experiments.

Auch die Stimulusauswahl sowie die Eingaben der Teilnehmerinnen und Teilnehmer sind Bestandteil der Sprachdatenbank.

3.6 Dokumentation

Der gesamte Prozess der Erstellung der *Brothers*-Sprachdatenbank wurde detailliert dokumentiert, u. a. mit Fotos der Aufnahmekabinen und genauer Angaben der bei den Aufnahmen verwendeten Geräte. Diese Dokumentation bildet ein eigenes Kapitel der Dissertation.

3.7 Validierung

Nach Abschluss der Dissertation wurde die *Brothers* Sprachdatenbank vor der Veröffentlichung im BAS Repository im Mai 2015 validiert. Der Validierungsbericht ist als Teil der Sprachdatenbank dort verfügbar.

Bei dieser Validierung wurde festgestellt, dass die Qualität der automatischen Segmentation für die Studioaufnahmen deutlich besser ist als für die Telefonaufnahmen. Gemessen wurde der prozentuale Anteil an falsch alignierten Wörtern im Text. Dieser Anteil lag bei 9 % für die Studioaufnahmen und 24 % für die Telefonaufnahmen, d. h. dass fast jedes vierte Wort im automatischen Verfahren nicht korrekt segmentiert wurde.

3.8 Verfügbarkeit der Datenbank

Die Datenbank *Brothers* ist über das Repository des BAS über den *persistent identifier* (PID) http://hdl.handle.net/11022/1009-0000-0001-55C3-3 für akademische Nutzer frei verfügbar, unter anderem im Emu-Datenbankformat.

4 Zusammenfassung

Brothers ist ein Beispiel für eine phonetische Sprachdatenbank, die zunächst konkret für die Untersuchung der Stimmähnlichkeit von Brüdern im Rahmen einer Dissertation entwickelt und ausgewertet wurde, und die danach so aufbereitet wurde, dass sie in das CLARIN Repository des Bayerischen Archivs für Sprachsignale aufgenommen werden konnte.

Im Repository stehen nun nicht nur die in der Dissertation ausgewerteten Roh- und abgeleiteten Messdaten, sondern auch noch weiteres Sprachmaterial wie z. B. die spontansprachlichen Dialoge, die noch nicht transkribiert oder ausgewertet wurden, sowie ein Validierungsbericht zur Sprachdatenbank mit Angaben zur Qualität der automatischen Segmentation.

Die wesentlichen Ergebnisse der Dissertation sind, dass

1. Brüderpaare in Perzeptionsexperimenten anhand der Stimme signifikant häufiger als solche erkannt wurden als Nichtbrüder,
2. die Stimmen von Brüderpaaren *auditiv* häufiger miteinander verwechselt werden als die Stimmen von nicht-verwandten Sprechern,
3. diese Verwechslung beim Telefon höher ist als bei Studioaufnahmen und dass
4. die *akustischen* Messungen für einzelne Sprecher jeweils charakteristisch sind, aber mit Ausnahme der Lesebedingung nicht zur Trennung von Brüderpaaren und nicht-verwandten Sprechern geeignet sind.

Feiser (2015) zieht daraus den Schluss, dass die perzeptive Ähnlichkeit von Brüderstimmen kaum von den akustischen Merkmalen abhängt, sondern „eher das erworbene Sprecherverhalten der Geschwister“ widerspiegelt (Feiser 2015: 140). Das Beispiel der Sprachdatenbank *Brothers* zeigt zum einen, dass Sprachdatenbanken eine unverzichtbare Grundlage der systematischen Untersuchung phonetischer Fragestellungen sind, zum anderen, dass damit Sprachressourcen, bestehend aus Primär-, Sekundär- und Metadaten in nachhaltiger Form vorgehalten und für weitere Untersuchungen genutzt werden können.

Literatur

Becker, Thomas (2012): *Einführung in die Phonetik und Phonologie des Deutschen*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Bird, Steven & Mark Liberman (2001): A formal framework for linguistic annotation. *Speech Communication* 33 (1,2), 23–60.

Boersma, Paul & David Weenink (1996): Praat, a system for doing phonetics by computer. Tech. Rep. 132 Institute of Phonetic Sciences of the University of Amsterdam.

Cassidy, Steve & Jonathan Harrington (1996): Emu: An enhanced hierarchical speech data management system. In *Proc. SST*, 361–366. Adelaide.

Cassidy, Steve & Jonathan Harrington (2001): Multi-level annotation in the emu speech database management system. *Speech Communication* 33, 61–77.

Draxler, Christoph (2008): *Korpusbasierte Sprachverarbeitung – eine Einführung*. Tübingen: Gunter Narr.

Draxler, Christoph (2011): Percy – An HTML5 framework for media rich web experiments on mobile devices. In *Proc. Interspeech*, 3339–3340. Florence, Italy.

Draxler, Christoph, Toomas Altosaar, Sadaoki Furui, Mark Liberman & Peter Wittenburg (2011): Speech processing tools – an introduction to interoperability. In *Proc. Interspeech*, 3229–3232. Florence, Italy.

Draxler, Christoph & Klaus Jänsch (2004): SpeechRecorder – a universal platform independent multi-channel audio recording software. In *Proc. LREC*, 559–562. Lisbon, Portugal.

Draxler, Christoph & Stefan Kleiner (2015): A cross-database comparison of two large german speech databases. In *Proceedings ICPhS*, Glasgow.

Feiser, Hanna (2015): *Untersuchung auditiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Brüderpaaren unter forensischen Aspekten*. Frankfurt am Main: Verlag für Polizeiwissenschaft.

Garofolo, John, Lori Lamel, William Fisher, Jonathan Fiscus, David S. Pallett & Nancy Dahlgren (1986): *The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM*. NIST.

Gibbon, Dafydd, Roger Moore & Richard Winski (1997): *Handbook of standards and resources for spoken language systems*. Berlin: Mouton de Gruyter.

Gibbon, Fiona & Lisa Crampin (2001): An electropalatographic investigation of middorsum palatal stops in an adult with repaired cleft palate. *Cleft Palate Craniofacial Journal* 38, 96–105.

Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (Hrsg.) (2006): *Essentials of language documentation*. Mouton de Gruyter.

Harrington, Jonathan (2010): *Phonetic analysis of speech corpora*. Oxford: Wiley-Blackwell.

Harrington, Jonathan, Steve Cassidy, Janet Fletcher & Andrew McVeigh (1993): The MU+ system for corpus based speech research. *Computer Speech and Language* 7, 305–331.

Holmqvist, Kenneth, Marcus Nyström & Fiona Mulvey (2012): Eye tracker data quality: what it is and how to measure it. *Proceedings acm Symposium on Eye Tracking Research and Applications* 45–52.

Huettig, Falk, Joost Rommers & Antje Meyer (2011): Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica* 137, 151–171.

IPA (1989): IPA Kiel Convention Workgroup 9 Report. *Journal of the IPA* 19 (2), 81–82.

IPA (1999): *Handbook of the IPA*. Cambridge: Cambridge University Press.

Lemnitzer, Lothar & Heike Zinsmeister (2006): *Korpuslinguistik – eine Einführung*. Tübingen: Narr Francke Attempto.

Narayanan, Shrikanth, Erik Bresch, Prasanta Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon Kim, Adam Lammert, Michael Proctor, Vikram Ramanarayanan & Yinghua Zhu (2011): A multimodal real-time MRI articulatory corpus for speech research. In *Proceedings Interspeech*, Florence.

Narayanan, Shrikanth, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis & Michael Proctor (2014): Real-

time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *Journal of the Acoustical Society of America* 136(3), 1307–1311.

Pömp, Julian & Christoph Draxler (2017): OCTRA – A configurable browser-based editor for orthographic transcription. In *Proceedings Phonetik und Phonologie*, Berlin.

Pompino-Marschall, Bernd (1995): *Einführung in die Phonetik*. Berlin: de Gruyter Mouton.

Reetz, Henning & Allard Longman (2009): *Phonetics – transcription, production, acoustics and perception*. Blackwell.

Richardson, Matt, Jeff Bilmes & Chris Diorio (2003): Hidden-articulator markov models for speech recognition. *Speech Communication* 41 (2–3), 511–529.

Richmond, Korin, Phil Hoole & Simon King (2011): Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech*, 1505–1508.

Schiel, Florian, Christoph Draxler, Angela Baumann, Tania Ellbogen & Alexander Steffen (2003): *The production of speech corpora*. Institut für Phonetik und Sprachliche Kommunikation, Universität München.

Schmidt, Thomas, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Dan Loehr, Magnus Magnusson, Travis Rose & Han Sloetjes (2009): An exchange format for multimodal annotations. In *Multimodal Corpora* (Lecture Notes in Computer Science 5509), 207–221. Springer.

Schmidt, Thomas & Kai Wörner (2005): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung* 6, 171–195.

Sloetjes, Han, Albert Russel & Alex Klassmann (2007): ELAN: a free and open-source multimedia annotation tool. In *Proc. Interspeech*, 4015–4016. Antwerp.

Stone, Maureen (2005): A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics* 19 (6–7), 455–501.

Westbury, John R., Greg Turner & Jim Dembrovski (1994): X-ray microbeam speech production database user's handbook. Tech. rep. Waisman Center, Washington University.

Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch (2017): Emu-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech and Language*.

Wrench, Alan A. & William J. Hardcastle (2000): A multichannel articulatory speech database and its application for automatic speech recognition. In *Proc. 5th Seminar on Speech Production*, 305–308.

Thomas Schmidt

9 Gesprächskorpora

Aktuelle Herausforderungen für einen besonderen Korpustyp

Abstract: Dieser Beitrag setzt sich mit Gesprächskorpora als einem besonderen Typus von Korpora gesprochener Sprache auseinander. Es werden zunächst wesentliche Eigenschaften solcher Korpora herausgearbeitet und einige der wichtigsten deutschsprachigen Gesprächskorpora vorgestellt. Der zweite Teil des Beitrags setzt sich dann mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) auseinander. FOLK hat sich zum Ziel gesetzt, ein wissenschaftsöffentliches Korpus von Interaktionsdaten aufzubauen, das methodisch und technisch dem aktuellen Forschungsstand entspricht. Die Herausforderungen, die sich beim Aufbau von FOLK in methodischer und korpustechnologischer Hinsicht stellen, werden in abschließenden Abschnitt diskutiert.

Keywords: Gesprächsforschung, gesprochene Sprache, Interaktion, Korpuslinguistik

1 Einleitung

Wenn gesprochene Daten in der Korpuslinguistik generell schon eine Sonderrolle einnehmen (siehe Mair in diesem Band), so stellen Gesprächskorpora noch einmal einen besonderen Fall unter den mündlichen Korpora dar, der ganz eigene Forschungsperspektiven und methodisch-technische Herausforderungen mit sich bringt. Um diese Perspektiven und Herausforderungen soll es im vorliegenden Beitrag gehen. Ich arbeite in Abschnitt 2 zunächst wesentliche Eigenschaften solcher Korpora heraus, die auch dazu dienen, sie von anderen Korpora mündlichen Sprachgebrauchs zu unterscheiden. Abschnitt 3 stellt dann einige der wichtigsten deutschsprachigen Korpora vor und geht auf die Problematik ein, dass ältere Sammlungen von Gesprächsdaten oft kaum für eine korpuslinguistische Nachnutzung zur Verfügung stehen. Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), das in Abschnitt 4 vorgestellt

Thomas Schmidt, Institut für Deutsche Sprache, R5, 6–13, D-68161 Mannheim,
E-Mail: thomas.schmidt@ids-mannheim.de

https://doi.org/10.1515/9783110538649-010

wird, hat sich vor diesem Hintergrund zum Ziel gesetzt, der Forschergemeinschaft ein großes, breit diversifiziertes Korpus von Interaktionsdaten zur Verfügung zu stellen. Aktuelle Herausforderungen, die sich beim Aufbau von FOLK in methodischer und korpustechnologischer Hinsicht stellen, werden dann in Abschnitt 5 diskutiert.

2 Gesprächskorpora als besondere mündliche Korpora

Unter einem Gesprächskorpus soll hier ein Korpus verstanden werden, das folgende Eigenschaften besitzt:

1. Die Primärdaten sind Audio- und/oder Videoaufzeichnungen von *Gesprächen*, also von verbaler Interaktion zwischen zwei oder mehr Teilnehmer(innen).
2. Die aufgezeichneten Gespräche sind *authentische* (natürliche) Interaktionen in dem Sinne, dass sie nicht eigens vom Forscher veranlasst wurden (wie es etwa bei einem Sprachexperiment, einem Interview oder laborphonetischen Daten der Fall ist).[1]
3. Die aufgezeichneten Äußerungen bestehen weitestgehend aus *spontanen* Äußerungen in dem Sinne, dass sie in ihrer konkreten Form nicht umfassend vorgeplant wurden (wie es etwa bei einer abgelesenen Rede oder einem geskripteten Dialog der Fall wäre).
4. Die Aufzeichnungen sind *vollständig*, erstens in dem Sinne, dass nicht nur ein zeitlicher Ausschnitt aus der Interaktion aufgezeichnet wird, zweitens auch in dem Sinne, dass die Aufzeichnung die Beiträge *aller* Interaktions-Teilnehmer(innen) gleichberechtigt umfasst (und nicht etwa durch die Aufnahmetechnik nur ausgewählte Teilnehmer fokussiert werden).

In der Gesamtheit dieser Eigenschaften unterscheiden sich Gesprächskorpora von anderen mündlichen Korpora, insbesondere von den meisten Variations-

---

1 Die Authentizität steht dabei immer in einem Spannungsverhältnis zum Beobachter-Paradoxon, das Labov (1972: 209) wie folgt beschreibt: „[T]he aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation.“ Die Gesprächsforschung hat sich mit diesem Paradoxon eingehend auseinandergesetzt und Methoden entwickelt, die negativen Auswirkungen des Paradoxons auf die Authentizität von Gesprächsdaten zu minimieren (z. B. Lalouschek & Menz 2002).

korpora (siehe Kehrein & Vorberger i. d. Bd. und Boas & Fingerhuth i. d. Bd.) und den meisten Sammlungen mündlicher Daten, die in der Phonetik oder Sprachtechnologie zum Einsatz kommen (siehe Draxler & Schiel in diesem Band). Die Grenzen zwischen einem Gesprächskorpus und anderen Typen mündlicher Korpora sind im Einzelfall allerdings nicht eindeutig zu ziehen: Einerseits weisen auch als Gesprächskorpora konzipierte Datensammlungen Merkmale auf, die es erlauben, variationslinguistische, phonetische oder sprachtechnologische Untersuchungen an ihnen auszuführen – beispielweise beinhaltet ein Korpus wie FOLK natürlich auch regionalsprachliche Variation. Andererseits lassen sich auch anders konzipierte Korpora teilweise unter gesprächsanalytischer Perspektive betrachten – beispielsweise können auch biographische oder narrative Interviews, wie sie für die dialektologische Tradition (etwa im Korpus „Deutsche Mundarten", siehe auch Kehrein & Vorberger in diesem Band) charakteristisch sind, als spezielle Formen von Gesprächen untersucht werden. Es geht hier daher nicht um eine absolute taxonomische Abgrenzung, sondern darum, den Begriff „Gesprächskorpus" für solche Korpora zu reservieren, bei deren Design, Umsetzung und Anwendung der Gedanke von Sprache im interaktiven Handeln leitend ist.

In ihrem Datenverständnis sind Gesprächskorpora vor allem von gesprächsanalytischen Forschungsansätzen wie der Konversationsanalyse, der Interaktionalen Linguistik oder der Funktionalen Pragmatik geprägt. Dies zieht zum einen eine besondere Aufmerksamkeit seitens der Korpusersteller und -nutzer für die sozialen Hintergründe und Zusammenhänge, denen die Daten entstammen und die sich in ihnen widerspiegeln, nach sich. Zum anderen ergibt sich daraus auch eine gewisse Zurückhaltung in der apriorischen Kategorisierung und „Kontrolle" von Variablen des Korpus-Designs, die sich einer korpuslinguistischen Methodik – zumindest in einer stärker quantitativen Orientierung – zunächst entgegenstellt. Da bislang weder die Gesprächsanalyse über ein fortgeschrittenes Instrumentarium zur Korpusanalyse verfügt noch die Korpuslinguistik sich umfassend mit dem Datentyp „Gespräch" auseinandergesetzt hat, erfordern der Aufbau und die Analyse von Gesprächskorpora also auch methodische Innovationen (siehe dazu auch Schmidt 2014a).

Zu den ganz konkreten methodischen Fragen, die sich für den Typus „Gesprächskorpus" exklusiv oder zumindest in besonderem Maße stellen, gehören:

- Welche Kontextinformationen oder Zusatzmaterialien sind zu einem möglichst vollständigen Verständnis eines Gesprächs und zu seiner Interpretation notwendig? Hierzu gehören zum einen Metadaten, die beispielsweise die institutionelle Einbettung eines Gesprächs oder die sozialen Rollen der Teilnehmer(innen) hinreichend genau beschreiben. Zum anderen können hier auch Objekte, die von der Aufnahme nicht (vollständig) erfasst wer-

den, eine Rolle spielen, wie z. B. PowerPoint-Folien bei einem mündlichen Vortrag, Zeichnungen oder Notizen, die im Laufe einer Besprechung angefertigt werden.
- Auf welche Art und Weise werden die Audio- oder Video-Daten auf eine schriftliche und damit automatisch durchsuchbare Form abgebildet? Die theorie- und forschungsfragenabhängige, modellhafte Relation zwischen den Primärdaten (den Aufnahmen) und ihrer Erschließung in Form von Transkription ist vielfach thematisiert worden (dazu Ochs 1979 und bspw. Schmidt 2005a). Sie spielt auch im Kontext korpuslinguistischer Herangehensweisen eine fundamentale Rolle, denn letztendlich bilden ja die schriftlichen Abbilder, nicht die Aufnahmen selbst, den Ausgangspunkt jeder Analyse. Die folgenden Fragen können auch als Teilaspekte dieser übergeordneten Frage aufgefasst werden (siehe auch Abb. 9.1, aus der deutlich wird, in welcher Dichte diese Fragen auftreten).
- In welcher Form sollen auch nonverbale Elemente der Interaktion berücksichtigt werden? Es geht dabei um hörbare (z. B. Räuspern, Lachen, Geräusche) als auch sichtbare (Mimik, Gestik, Handlungen) Bestandteile der Aufzeichnungen, die hinsichtlich ihrer kommunikativen Relevanz beurteilt und mit geeigneten, idealerweise auch konsistent recherchierbaren, Beschreibungen versehen werden müssen.
- Wie soll die zeitliche Struktur sprachlicher Interaktionen repräsentiert werden? Hierbei geht es zum einen generell um die Frage, wie (z. B. in welcher Granularität) zeitliche Bezüge zwischen Transkription und Aufnahme („Text-Ton-Alignment") in den Daten festgehalten werden, zum anderen ganz besonders auch um den Umgang mit zeitlicher Parallelität, die in authentischen Gesprächen allgegenwärtig ist („Überlappungen", „Back channelling").
- Wie sollen Phänomene mündlicher „Performanz" wie Pausen, Häsitationen, Verschleifungen, Elisionen, Abbrüche oder Reparaturen, insbesondere auch in ihrer Relevanz für die Interaktion, in den Daten festgehalten und bei Analysen berücksichtigt werden? Wie ist mit dem Umstand umzugehen, dass die Untersuchung von authentischer gesprochener Sprache in der Regel nur unter nicht-optimalen akustischen Bedingungen erfolgen kann? Hintergrundgeräusche, schwankende Aufnahmequalität u. Ä. und die daraus resultierende schwere Verständlichkeit mancher Äußerungen können dazu führen, dass einige Teile der Interaktion weniger genau und zuverlässig beschrieben werden können als andere.

Dass diese Aspekte Gesprächskorpora zu einem besonderen Korpustyp machen, wird deutlich, wenn man ihnen den Prototyp eines schriftsprachlichen Korpus

0011	**AM**	was
0012	**PB**	hier bitteschön
0013	**AM**	oh (.) danke schon zucker rein getan
0014	**PB**	nee aber ich hab den dir hier mitgebracht
0015		(0.72)
0016	PB	das_s ungesüßt das schme[ckt dir]
0017	**AM**	[das is aber kein] latte macchiato
0018		(0.67)
0019	**PB**	[((lacht))] das_s cappuccinopulver mit viel milch (.) das wird (dafür/das ja) ausreichen ((Störgeräusch))
0020	**AM**	[des is n cappuccino]
0021		(0.58)
0022	AM	würdst du das bitte da liegen lassen

Abb. 9.1: Transkriptausschnitt aus FOLK_E_00043_SE_01_T_01 in der DGD (Alltagsgespräch – Paargespräch), mit Beispielen für die Transkription von Pausen (Zeilen 13, 15, 18, 21), Überlappungen (Zeilen 16/17 und 19/20), Verschleifungen (Zeilen 16 und 19), Elisionen (Zeilen 20 und 22), Unsicherheit des Transkribenten (Zeile 19).

redigierter, veröffentlichter Texte gegenüberstellt. Dieser „Default Case“ zeichnet sich nämlich gerade dadurch aus, dass die darin enthaltene Sprache weitestgehend losgelöst von ihrem außersprachlichen Kontext (als „Sprache der Distanz“ nach Koch & Oesterreicher 1985), frei von äußeren „Störungen“ und als einfache lineare Abfolge eindeutiger, nicht interpretationsbedürftiger sprachlicher Zeichen analysiert werden kann, die genannten Aspekte dort mithin kaum eine Rolle spielen. Dementsprechend können Methoden, Werkzeuge und Datenmodelle, die für schriftsprachliche Korpora entwickelt und angewendet werden, in der Regel nicht direkt auf die Arbeit mit Gesprächskorpora übertragen werden, sondern es sind hierfür eigene Arbeitsabläufe notwendig (siehe z. B. Schmidt 2016b).

3 Gesprächskorpora des Deutschen

Der Ausrichtung dieses Bandes folgend beschränke ich mich hier weitestgehend auf Gesprächskorpora des Deutschen. Der Vollständigkeit halber sei jedoch eingangs festgestellt, dass auch für viele andere Sprachen Datensammlungen existieren, die nach den oben genannten Kriterien als Gesprächskorpora zu klassifizieren sind. Für das amerikanische Englisch können das *Newport Beach Corpus* (Jefferson) und das *Santa Barbara Corpus of Spoken American English* (Du Bois et al. 2000–2005), für das australische Englisch das *Griffith Corpus of Spoken Australian* English (Haugh & Chang 2013) als prototypische Vertreter gelten. Das *British National Corpus* (BNC) hat insbesondere der Erhebung des „context-governed part“ seiner „spoken component“ eine

Systematik zugrunde gelegt, aufgrund derer man das resultierende Teilkorpus als Gesprächskorpus betrachten kann. Ähnliches gilt für Teile der Neuerhebungen im *Spoken BNC2014* (Love et al. 2017) oder auch für die betreffenden Subkorpora des *Corpus Gesproken Nederlands* (CGN, Oostdijk 2002), wobei letzteren drei Korpora gemein ist, dass sie gerade nicht als Gesprächskorpora konzipiert sind, sondern Gesprächsdaten lediglich als ein Datentypus unter anderen in das Design des Gesamtkorpus aufgenommen wurden. Für das Französische weisen das *Corpus International Écologique de la Langue Française* (CIEL-F, Dister et al. 2008) wie auch mehrere Korpora, die in der Plattform *Corpus de Langue Parlé en Interaction* (CLAPI, Groupe ICOR 2010) verfügbar sind, recht eindeutig die wesentlichen Eigenschaften eines Gesprächskorpus auf, und auch beim Aufbau des ESLO-Korpus (Eshkol-Taravella et al. 2012) werden nach anfänglicher Konzentration auf soziolinguistische Interviews neuerdings vermehrt authentische Interaktionsdaten berücksichtigt.

Für das Deutsche kann das Korpus „Grundstrukturen“ – oft besser bekannt unter dem Namen „Freiburger Korpus“ –, das als „Korpus der alltäglichen, übergruppal und überregional verstandenen und akzeptierten gesprochenen deutschen Standardsprache“ (vgl. Schröder 1975: 12) konzipiert wurde, als Wegbereiter für das Feld der Gesprächskorpora gelten. In einem Folgeprojekt wurde mit dem Korpus „Dialogstrukturen“ (Berens et al. 1976) eine weitere Datensammlung aufgebaut, die noch dezidierter auf den Interaktionsaspekt mündlichen Sprachgebrauchs fokussiert ist. Beide Korpora entstanden in Projekten des Instituts für Deutsche Sprache (IDS) und stehen am Anfang einer Hinwendung des Instituts zu soziolinguistisch und pragmatisch ausgerichteter Forschung, die in den 1980er und 1990er Jahren ihren Niederschlag auch im Aufbau weiterer Gesprächskorpora fand. Zu nennen sind hier mindestens die Korpora „Stadtsprache Mannheim“ (Kallmeyer 1994), „Beratungsgespräche“ (Nothdurft, Reitemeier & Schröder 1994), „Gespräche im Fernsehen“ (Schütte 1996), „Schlichtungs- und Gerichtsverhandlungen“ (Schröder 1997) und „Deutsch-Türkische Powergirls“ (Keim 2008), die sich jeweils als Gesprächskorpora verstehen lassen, in denen ein bestimmter Lebensraum, ein bestimmtes soziales Milieu bzw. ein bestimmter Interaktionstyp fokussiert wird.

Auch im Kontext funktional-pragmatischer Diskursanalysen entstanden in dieser Zeit größere Datensammlungen, die sich als Gesprächskorpora – hier oft mit Ausrichtung auf einen bestimmten institutionellen Kontext – qualifizieren lassen, z. B. zur „Analyse von Unterrichtskommunikation“ (Ehlich & Rehbein 1986), zur „Ausbildung im Bergbau“ (Brünner 1987) oder zu „Sprachlichen Verständigungsprozessen in der Arzt-Patienten-Kommunikation“ (Rehbein & Löning 1993), später auch mit einem speziellen Blick auf Mehrsprachigkeit wie in „Die Entwicklung narrativer Diskursfähigkeiten im Deutschen und Tür-

kischen in Familie und Schule“ (ENDFAS/SKOBI, Herkenrath & Rehbein 2012), zur „Sprache der Höflichkeit in der interkulturellen Kommunikation“ (SHiK, Rehbein et al. 2001), zu „Japanischen und deutschen Expertendiskursen in ein- und mehrsprachigen Konstellationen“ (JadEx, Hohenstein 2006) oder zum „Dolmetschen im Krankenhaus“ (DiK, Bührig et al. 2012).

An jüngeren Initiativen, die den Aufbau von Gesprächskorpora zum Gegenstand haben, sind insbesondere das „Kiezdeutschkorpus“ (KiDKo, Wiese et al. 2012) und das Korpus „Gesprochene Wissenschaftssprache Kontrastiv“ (GeWiss, Fandrych, Meißner & Slavcheva 2012) erwähnenswert, außerdem die Datenbank „Gesprochenes Deutsch für die Auslandsgermanistik“ (Imo & Weidner in diesem Band).

Mit den bislang genannten Korpora ist allerdings nur ein kleiner Teil der Gesprächsdaten erfasst, die seit der „pragmatischen Wende“ in Forschungsprojekten erhoben wurden und theoretisch für Korpusanalysen verwendet werden könnten. Ein erschöpfender Überblick scheint zum einen mangels vollständiger Informationsquellen kaum möglich (siehe aber Wagener & Bausch 1997 und Glas & Ehlich 2000). Zum anderen offenbart schon ein zweiter Blick auf die genannten Korpora das grundlegende Problem, dass viele Gesprächskorpora, die als empirische Grundlage für Einzelprojekte aufwändig erhoben und erschlossen wurden, nach Abschluss dieser Projekte oft nicht für weitere Analysen zur Verfügung stehen.[2] Die Gründe hierfür sind vielfältig: neben ungelösten technischen Herausforderungen und dem Mangel an Bereitschaft zum Teilen von Daten (siehe dazu auch Schmidt 2005b) fehlt bei vielen Korpora die rechtliche Grundlage für eine Weitergabe der Daten, weil bei ihrer Erhebung keine geeignete Einwilligung der Gesprächsteilnehmer eingeholt wurde. Dies ist zum Teil dem Umstand geschuldet, dass rechtliche Fragen einer Datenweitergabe bei der Erhebung nicht oder nicht ausreichend bedacht wurden bzw. – mit Blick auf die Distribution von Daten in elektronischer Form über das

---

2 Die Korpora „Grundstrukturen“ und „Dialogstrukturen“ sind vollständig über das Archiv für Gesprochenes Deutsch (AGD) und die Datenbank für Gesprochenes Deutsch (DGD) verfügbar, die weiteren im IDS-Kontext entstandenen Korpora jedoch nur in kleineren Auszügen oder gar nicht. Das Korpus „Ausbildung im Bergbau“ wurde 2014 ins AGD integriert, die Audio- und Videoaufnahmen sind über den persönlichen Archivservice erhältlich. Die Korpora ENDFAS/SKOBI und DiK werden über das Hamburger Zentrum für Sprachkorpora (https://corpora.uni-hamburg.de/hzsk/ [letzter Zugriff: 26. 9. 2017]) archiviert und weitergegeben, allerdings ohne das zugrunde liegende Audio. Gleiches gilt für das KiDKo, das über die ANNIS-Plattform an der HU Berlin (https://korpling.german.hu-berlin.de/annis3/ [letzter Zugriff: 26. 9. 2017]) verfügbar gemacht wird. Für das GeWiss-Korpus wird an der Universität Leipzig eine eigene Zugriffs-Plattform (https://gewiss.uni-leipzig.de/ [letzter Zugriff: 26. 9. 2017]) betrieben, zusätzlich wird das Korpus aktuell in die Bestände des AGD integriert. Für alle anderen genannten Korpora müssen Bedingungen der Archivierung und möglichen Weitergabe aktuell als ungeklärt gelten.

WWW – gar nicht bedacht werden konnten. Darüber hinaus fragt sich jedoch, ob die betreffenden Datensammlungen bei ihrer Entstehung überhaupt als „Korpora“ konzipiert worden waren: die Erwartung, dass sie als empirische Datengrundlage über den eigentlichen Projektkontext hinaus verwendet werden können oder gar sollten, mag aus heutiger Perspektive selbstverständlich erscheinen; für ein ethnographisch oder soziolinguistisch orientiertes Forschungsprojekt in den 1970er oder 1980er Jahren kann sich dies jedoch grundlegend anders dargestellt haben.

4 FOLK

Obwohl die empirische Arbeit mit Gesprächsdaten also in Deutschland auf eine mehrere Jahrzehnte umfassende Tradition zurückblickt, standen der germanistischen Sprachwissenschaft noch zu Beginn des Jahrtausends kaum allgemein nutzbare Gesprächskorpora zur Verfügung. Dieser Missstand war für das IDS der Anlass, ein Projekt zum Aufbau eines „nationalen Gesprächskorpus“ (Deppermann & Hartung 2012) zu initiieren:

> Aufgrund dieser unbefriedigenden Situation haben wir im Jahre 2008 am IDS damit begonnen, ein nationales Gesprächskorpus aufzubauen, das den „kommunikativen Haushalt“ (Luckmann 1986) der deutschsprachigen mündlichen Kommunikationspraxis in seinen wesentlichen Ausprägungen repräsentieren soll [...]. Das regulative Ziel ist es, das volle Spektrum der privaten, institutionellen, öffentlichen und massenmedialen Anlässe und Typen mündlicher Kommunikation nach und nach durch Audio- und Videoaufnahmen zu dokumentieren, zu transkribieren und soweit als möglich der wissenschaftlichen Gemeinschaft zur Verwendung für Forschungs- und Lehrzwecke zur Verfügung zu stellen. Dementsprechend nennen wir dieses Korpus „Forschungs- und Lehrkorpus gesprochenes Deutsch“ (FOLK [...]). (Deppermann & Hartung 2012: 418)

FOLK wurde Ende 2012 erstmalig mit Version 2.0 der *Datenbank für Gesprochenes Deutsch* (DGD, Schmidt 2014b) veröffentlicht und wird seitdem kontinuierlich ausgebaut. Die aktuelle Version (vom April 2017) umfasst 259 Gespräche im Umfang von etwas mehr als 202 Stunden und ca. 2 Millionen transkribierten Tokens, die verschiedenste Interaktionstypen aus den Bereichen privater (z. B. Tischgespräche, Telefongespräche, Spielinteraktionen, Gespräche bei privaten Aktivitäten), institutioneller (z. B. schulischer Unterricht, Verkaufsgespräche, Fahrschulstunden, berufliche Gespräche, universitäre Prüfungsgespräche) und öffentlicher Kommunikation (z. B. Podiumsdiskussion, Schlichtungsgespräch) abdecken.

Die korpustechnologischen Werkzeuge und Verfahren, die beim Aufbau von FOLK zum Einsatz kommen und zu einem nicht unerheblichen Teil eigens

für dieses Projekt entwickelt oder optimiert wurden, sind in Schmidt (2016b) näher beschrieben. Sie dienen alle dem übergeordneten Ziel, die Erstellung eines Gesprächskorpus nicht nur praktisch und technisch handhabbar zu machen, sondern die entstehenden Daten darüber hinaus anschlussfähig an gute Praktiken (dazu Schmidt 2016a) und anderweitig etablierte digitale Verfahren in der Korpuslinguistik zu machen. Somit soll sich FOLK als Gesprächskorpus bei allen Besonderheiten und Unterschieden zum „Default Case" des schriftsprachlichen Korpus (siehe dazu auch Kupietz & Schmidt 2015) mittelfristig in ein Gesamtgefüge einordnen, in dem auch gemeinsame oder kontrastierende Untersuchungen über verschiedene Korpustypen hinweg ermöglicht werden.

Wie aus mittlerweile über 7.000 Registrierungen für die DGD, in der FOLK das mit Abstand am meisten genutzte Korpus ist, ersichtlich ist, wird mit der Bereitstellung dieses Gesprächskorpus ein realer und großer Bedarf von Forschenden, Lehrenden und Studierenden adressiert. Eine systematische Nutzerstudie (Fandrych et al. 2016) hat gezeigt, dass die Daten in unterschiedlichsten Anwendungsszenarien zum Einsatz kommen. FOLK wird demnach außer als Basis-Ressource in der sprachwissenschaftlichen universitären Ausbildung und als Datengrundlage für gesprächsanalytische Arbeiten insbesondere auch für variationslinguistische Untersuchungen, für vergleichende Korpusanalysen, als Ressource für die Sprachvermittlung im Bereich DaF/DaZ und als Quelle für die Entwicklung sprachtechnologischer Anwendungen fruchtbar gemacht. Auffällig ist darüber hinaus ein großes Interesse an FOLK in der Auslandsgermanistik, das sich nicht zuletzt dadurch erklären lässt, dass mit FOLK Studierenden im Ausland ein einfacher Zugriff auf authentische und aktuelle Gesprächsdaten des Deutschen in größerer Vielfalt ermöglicht wird (siehe dazu auch Imo & Weidner in diesem Band).

Exemplarische Analysen, die anhand von FOLK gesprächsanalytische Methodik mit korpuslinguistischen Verfahren kombinieren, finden sich beispielsweise in Deppermann & Schmidt (2014), Schmidt (2014a) und Kaiser (2017), wo jeweils einzelne Diskursmarker (*das heißt*, *ich sag mal*, bzw. *sprich*) untersucht werden. Mehrere aus dem Projekt „Verbkomplemente im gesprochenen Deutsch" hervorgegangene Arbeiten (vgl. Deppermann et al. 2017) widmen sich der deskriptiven und funktionalen Beschreibung von Argumentstrukturen und Verbkomplementen und stützen sich sowohl bei der Kontrastierung von Mündlichkeit und Schriftlichkeit als auch bei der Untersuchung interaktionsspezifischer Besonderheiten wesentlich auf Korpus-Evidenz aus FOLK. In ähnlicher Weise hat das Projekt „Lexik des Gesprochenen Deutsch" (LeGeDe, Meliss & Möhrs 2017) ausgehend von FOLK begonnen, erstmalig den Wortschatz des gesprochenen Deutsch in der Interaktion mit korpuslexikographischen Methoden zu untersuchen und zu beschreiben.

5 Aktuelle Herausforderungen

Die bisherigen Arbeiten mit Daten aus FOLK machen bereits das große Potenzial deutlich, das ein wissenschaftsöffentlich verfügbares Gesprächskorpus für die sprachwissenschaftliche Forschung und Lehre birgt. Beim Aufbau von FOLK offenbaren sich aber auch eine ganze Reihe von Herausforderungen, die dieser spezielle Korpustyp mit sich bringt und deren Bearbeitung noch lange nicht als abgeschlossen betrachtet werden kann. Diese sind zum Teil eher theoretisch-methodischer, zum Teil eher praktisch-technologischer Natur, sie interagieren aber auch vielfältig miteinander. Exemplarisch seien im Folgenden die Bereiche des Korpus-Designs und der Korpustechnologie diskutiert.

5.1 Korpus-Design und Stratifikation

Als Referenzkorpus muss FOLK anstreben, seinen Gegenstand – Gesprächsinteraktionen im Deutschen – in möglichst großer Breite und Differenziertheit und nach einer nachvollziehbaren Systematik abzubilden. Leitend für das Korpus-Design ist dabei zunächst der Begriff des Gesprächstyps, d. h. vor allen anderen Eigenschaften sind es Unterschiede in Interaktionsanlässen, -konstellationen, -kontexten und -inhalten (i. w. S. „Situational Parameters" nach Biber 1993: 245), die im Korpus angemessen abgebildet werden müssen. Bei schriftsprachlichen Korpora können zumindest allgemeinere Kategorien wie „Zeitungstext", „Belletristik", „Gebrauchstext", „wissenschaftlicher Text" (vgl. das „Kernkorpus" des Digitalen Wörterbuchs der Deutschen Sprache, DWDS)[3] insofern als robust und etabliert gelten, als sie in dieser oder ähnlicher Benennung und Systematik beim Design mehrerer Referenzkorpora zur Anwendung kommen. Für die Binnendifferenzierung solcher übergeordneten Kategorien stehen außerdem oft zusätzliche externe Systematiken (wie Fachsystematik für wissenschaftliche Texte, Ressorts für Zeitungstexte, literarische Gattungen für Belletristik) zur Verfügung, die für eine detaillierte Korpus-Stratifikation fruchtbar gemacht werden können. Für die Klassifizierung mündlicher Interaktionen existiert keine vergleichbar stabile Ausgangslage. Als weitestgehend einfach operationalisierbar kann allenfalls eine erste Unterscheidung in Interaktionsdomänen gelten, die ein gegebenes Gespräch z. B. dem privaten, dem institutionellen oder dem öffentlichen Bereich (so in FOLK)[4]

---

3 https://www.dwds.de/ [letzter Zugriff: 07. 11. 2017].

4 Ähnlich z. B. bei Biber (1993: 245), wo unter dem Stichwort „Setting" zwischen „Institutional, other public, private-personal" unterschieden wird, oder beim slowenischen GOS-Korpus (Verdonik et al. 2013), dessen Bestandteile jeweils einer der drei Kategorien „Public, Non-pub-

zuordnet. Für eine weitere Binnendifferenzierung können bei institutionellen Gesprächen ggf. noch die betreffenden Institutionen selbst (z. B. Schule, Universität, Verein, Kirche usw.) und eventuell diesen eigene (quasi „institutionalisierte") Typisierungen (z. B. nach Schulfach oder Klassenstufe in der Schule, Seminare/Übungen vs. Prüfungen an der Universität, Vorstandssitzung vs. Mitgliederversammlung im Verein) herangezogen werden; insbesondere im privaten Bereich ist die eindeutige Zuordnung eines gegebenen Gesprächs innerhalb einer eindeutigen Typen-Hierarchie aber oft nicht möglich, gerade weil sich private Alltagsgespräche dadurch auszeichnen, dass ihre Form nicht oder nur in geringem Maße äußerlich vorgegeben ist. Deppermann & Hartung (2012: 423 f.) schlagen daher für Gesprächskorpora eine „parametrisierte Systematik" vor, die ein Gesprächsereignis statt durch eine einfache Zuordnung zu einem Typ durch ein Bündel von Merkmalen in Form von Attribut-Wert-Paaren charakterisiert. Angeführt werden z. B. Parameter wie „Teilnehmerzahl" (z. B. mit Werten „dyadisch" vs. „Mehrpersonengespräch"), „Vertrautheit der Teilnehmer" („unbekannt", „bekannt", „vertraut"), „Publikum" („ja", „nein") oder „Zugang" („geschlossen", „halb-öffentlich", „öffentlich"). Im Hinblick auf Korpus-Design und -Ausbau (aber auch bei der Analyse) ist ein solcher Ansatz prinzipiell praktikabler, weil er im Gegensatz zu einer fixierten Gattungssystematik weniger (evtl. theoretisch strittige) Festlegungen erfordert und auch Raum lässt, Gesprächsereignisse ins Korpus-Design zu integrieren, die bei der Planung nicht vorhergesehen wurden. Allerdings ist es alles andere als trivial, eine solche Systematik für die Anwendung auf reale Gesprächsaufnahmen zu operationalisieren, denn sie erfordert u. a. die Klärung von Grenzfällen (z. B. „Ist ein Prüfungsgespräch zwischen Student und Prüfer, bei dem ein Beisitzer anwesend ist, der aber nicht aktiv am Gespräch teilnimmt, dyadisch oder ein Mehrpersonengespräch?") und Definitionen oder Leitlinien für interpretative Entscheidungen (z. B. „Ab wann gelten Gesprächsteilnehmer als vertraut und nicht mehr als nur bekannt?"). Letztendlich bieten sich für die Operationalisierung daher korpuslinguistische Methoden an, in denen diese Parameter als globale Annotationen verstanden werden, deren Intersubjektivität durch explizite Leitlinien und Inter-Rater-Agreement-Messungen abgesichert werden kann. Somit können Design und Stratifikation eines Gesprächskorpus wie FOLK also nicht vollständig à priori „am Reißbrett" erfolgen, sondern müssen begleitend zum Aufbau anhand des jeweils schon vorliegenden Materials em-

---

lic non-private, Private" zugeordnet sind. Das CGN unterscheidet hingegen zunächst nur zwischen „Private" und „Public", beim „context-governed part" des BNC kommt eine Kategorie „Public/Institutional" zur Anwendung, die mit den Kategorien „Educational/Informative", „Business" und „Leisure" kontrastiert.

pirisch entwickelt und fortwährend verifiziert werden. Nachdem FOLK in einer ersten Phase zunächst überwiegend opportunistisch (also mit Aufnahmen leicht erreichbarer Gesprächsereignisse) aufgebaut und in einer zweiten Phase mit Blick auf eine möglichst breite Streuung über (vorläufig angenommene) Gesprächstypen ausgebaut wurde, ist mit den nun vorliegenden 259 Gesprächsereignissen eine Basis gegeben, um eine parametrisierte Systematik in dieser Weise korpuslinguistisch zu fundieren, d. h. an konkretem empirischen Material zu entwickeln und zu erproben.

Neben dem Gesprächstyp sind für das Design und die Stratifikation eines „nationalen Gesprächskorpus" jedoch auch demographische Kriterien, also Eigenschaften der aufgenommenen Sprecher, relevant. Als Mindestanforderung für FOLK kann in dieser Hinsicht gelten, dass weibliche und männliche Sprecher in vergleichbaren Mengen berücksichtigt werden, dass das Korpus regionale Variation in ausreichendem Maße abbildet und dass auch bezüglich Alter und Bildungshintergrund der Sprecher eine Ausgewogenheit – oder zumindest vollständige Abdeckung – angestrebt wird. Diese „sekundären" Stratifikationsparameter sind zwar in der Theorie einfacher zu handhaben als die zuvor genannten, da sie weitestgehend objektiv feststellbar (und in diesem Sinne „echte" Metadaten, keine Annotationen) sind. Allerdings stellt sich bei der praktischen Umsetzung das Problem einer kombinatorischen Explosion: Ein Korpus-Design, das alle dann festgelegten Parameter miteinander kreuzt und für jede mögliche Kombination eine Mindestmenge an Daten vorsieht (also z. B. mindestens eine Aufnahme eines dyadischen, geschlossenen Gesprächs zwischen einander vertrauten Teilnehmern aus dem norddeutschen Sprachraum mit männlichen Sprechern unter 30 Jahren mit höherem Bildungsabschluss, und desgl. für alle anderen Kombinationen), führt zwangsläufig zu Datenmengen, die praktisch nicht mehr erheb- und verarbeitbar sind. Für Design und Stratifikation von FOLK muss also ein Kompromiss gefunden werden, der demographische Kriterien nicht ignoriert, sich aber am organisatorisch Machbaren orientiert. Der aktuell in FOLK (und teilweise auch in anderen Gesprächskorpora, die sich dieser Frage stellen) favorisierte Ansatz sieht vor, zum einen die Zahl der Attribut-Wert-Kombinationen für die demographische Stratifikation möglichst gering zu halten (indem z. B. nur zwei oder drei Altersspannen oder nur vier bis sechs sprachliche Großregionen unterschieden werden), zum anderen eine systematische Streuung über solche Parameter nur für ausgewählte, möglichst alltägliche Gesprächstypen (wie privates Telefongespräch, Tischgespräch, berufliches Meeting) anzustreben.

Kompromisse bei Korpus-Design und Stratifikation werden aber nicht nur auf Grund begrenzter Kapazitäten für Datenerhebung und -verarbeitung notwendig; nicht wenige Gesprächsereignisse sind auch wegen schwieriger akus-

tischer Bedingungen (z. B. Gespräch in der Disko) oder mangelnder Vorhersehbarkeit (z. B. Unterhaltung bei einer zufälligen Begegnung) kaum erhebbar oder aufgrund erhöhter Sensibilität (z. B. psychotherapeutisches Gespräch) für ein wissenschaftsöffentliches Korpus nicht autorisierbar. Der Anspruch, in einem Gesprächskorpus „das volle Spektrum [von Gesprächen] zu dokumentieren" (Deppermann & Hartung 2012: 418) kann daher nur als ein Ideal verstanden werden, dem man sich bestenfalls soweit annähern kann, dass die Variation der Stratifikationsparameter in einer für den Korpusnutzer nachvollziehbaren Weise maximiert wird. Da in FOLK in diesem Sinne das Prinzip „Breite vor Tiefe" angewendet wird – der Aufnahme von Gesprächen mit bislang nicht besetzten Parameterkombinationen also üblicherweise der Vorzug vor der Erhebung weiterer Instanzen bereits vorhandener Typen gegeben wird –, ist das Korpus dann auch weniger zur Bearbeitungen solcher Fragestellungen geeignet, die ganz spezielle Interaktionspraktiken oder Sprechertypen in den Blick nehmen. FOLK versteht sich auch in dieser Hinsicht als ein Referenzkorpus, das als Vergleichsbasis für vorhandene (z. B. zur Hochschulkommunikation wie in GeWiss, zur Arzt-Patienten-Kommunikation wie in DiK, zur Kommunikation zwischen multi-ethnischen Sprecherinnen wie in KiDKo) oder als Orientierungspunkt für zukünftig zu erstellende spezialisierte Gesprächskorpora dienen kann. Dies gilt auch für Daten, die in FOLK aus organisatorischen oder prinzipiellen Erwägungen bis auf Weiteres unberücksichtigt bleiben werden, wie insbesondere Gesprächsdaten aus Österreich und der deutschsprachigen Schweiz und mehrsprachigen Kommunikationssituationen.[5]

5.2 Korpustechnologie

Wie oben angesprochen, können korpuslinguistische Verfahren zur Annotation und Analyse generell nicht einfach vom schriftsprachlichen auf den mündlichen Fall übertragen werden, und Gesprächsdaten weisen in dieser Hinsicht gegenüber anderen gesprochensprachlichen Daten noch einmal besonders komplexe Eigenschaften auf.

Aus korpustechnologischer Sicht zentral ist zunächst die Tatsache, dass die Erhebung und Grunderschließung von Gesprächsdaten kaum durch auto-

---

5 Deren offensichtliche Relevanz soll damit in keiner Weise in Frage gestellt werden – ihre Berücksichtigung würde aber die Komplexität des Korpus-Aufbaus um zusätzliche Dimensionen erweitern. Gesprächsdaten des Österreichischen werden in einigen Teilprojekten des Spezialforschungsbereichs „Deutsch in Österreich" erhoben. Der Aufbau eines Referenzkorpus zu Gesprächen in mehrsprachigen Konstellationen unter Beteiligung des Deutschen bleibt ein Desiderat, siehe dazu aber mehrere Beiträge in Schmidt & Wörner (2012).

matische sprach- oder texttechnologische Verfahren unterstützt werden kann (wohingegen die Akquise schriftsprachlicher Daten über geeignete Harvesting-Methoden oft fast vollständig automatisiert ist). Die Erhebung einer Gesprächsaufnahme erfordert einen geeigneten Feldzugang, der in aller Regel nur über persönliche Kontakte herzustellen ist, und auch die Aufnahme selbst muss i. d. R. von einer technisch verständigen Person vorbereitet, den Bedingungen der jeweiligen Situation angepasst und durchgeführt werden.[6] Für die anschließende Basiserschließung, d. h. Transkription, einer Aufnahme ist prinzipiell der Einsatz von Spracherkennungstechnologie denkbar und wurde exemplarisch auch schon erprobt (siehe z. B. Moore 2015). Erste Experimente in FOLK haben aber ergeben, dass beim dort vorliegenden Material in aller Regel (d. h. von wenigen Ausnahmen abgesehen) nur Worterkennungsraten von deutlich unter 50 % erreicht werden, so dass der resultierende Korrekturbedarf letztendlich mindestens den gleichen Aufwand verursacht wie eine rein manuelle Transkription. Ähnliches gilt für einfachere sprachtechnologische Verfahren wie „silence detection" (die Erkennung von Pausen zur Vorsegmentierung einer Aufnahme) oder „speaker diarization" (die Erkennung von Sprechern und Sprecherwechseln in der Aufnahme), die prinzipiell geeignet erscheinen, den manuellen Transkriptionsaufwand deutlich zu reduzieren, aber in der Anwendung auf Gesprächsdaten[7] zu fehlerhaft sind, um dieses Potenzial zu realisieren. Die Überwindung des „Transkriptionsflaschenhalses" („transcription bottleneck", Brinckmann 2009) mittels Sprachtechnologie bleibt daher bis auf Weiteres ein Wunschtraum.

Liegt zu einer Aufnahme erst einmal eine Transkription vor, ist es möglich, diese mit Annotationsverfahren, die ursprünglich für schriftsprachliche Daten entwickelt wurden, automatisch anzureichern. Im Falle von FOLK umfasst dies derzeit eine orthographische Normalisierung (also die Abbildung literarisch transkribierter Formen wie *zwohunnert* auf ihre standardorthographische Entsprechung *zweihundert*) und, darauf aufbauend, eine Lemmatisierung und ein

6 Ein gewisses Potenzial zur Zentralisierung (wenn auch nicht Automatisierung) besteht immerhin bei medial vermittelten Gesprächen, also z. B. Telefon- oder Skypegesprächen, die über geeignete Software aufgezeichnet werden können, ohne dass eine Anwesenheit des Forschers „vor Ort" notwendig wäre. Zentral akquiriert werden können außerdem Aufnahmen aus Rundfunk und Fernsehen, sofern geeignete Abmachungen mit den Sendeanstalten vorliegen.

7 Dies gilt nicht unbedingt in gleichem Maße für andere Typen mündlicher Daten. Wo immer die Aufnahmebedingungen soweit kontrolliert werden können, dass durchgängig hochwertige Audiodaten mit gleichen technischen Parametern und ohne größere Störungen entstehen, erhöhen sich die Erfolgsaussichten beim Einsatz von Spracherkennungstechnologie. Am AGD wird solche Technologie daher zunächst – und teilweise bereits erfolgreich – in der Anwendung auf Variationskorpora erprobt.

Part-of-Speech-Tagging. Die einzelnen Verfahren sollen hier nicht im Detail diskutiert werden (siehe dazu Westpfahl & Schmidt 2016 und Schmidt 2016b). Wichtig ist, dass sie zwar einerseits (teil-)automatisiert werden können, andererseits aber erst dann zufriedenstellende Ergebnisse liefern, wenn sie – mit nicht unerheblichem Aufwand – an die Eigenheiten mündlicher Daten im Allgemeinen und von Gesprächsdaten im Besonderen angepasst wurden. Für FOLK und die genannten Annotationstypen ist diese Anpassung mittlerweile erfolgt, für andere automatische Annotationsverfahren, die für die korpuslinguistische Analyse schriftsprachlicher Daten fruchtbar gemacht werden (z. B. Parsing, morphologische Annotation) steht sie noch aus.

Ähnliches gilt für Verfahren der automatisierten Auswertung, die sich in der Korpuslinguistik etabliert haben, beispielsweise Kookkurrenzprofile oder Kollokationsmaße. Exemplarisch zeigen etwa Batinic & Schmidt (2017) am Beispiel der Rekonstruktion separabler Partikelverben, dass automatisierte Verfahren nicht ohne Modifikation vom schriftsprachlichen auf den mündlichen Fall übertragen werden können, z. B. weil diesen Verfahren die vermeintliche Selbstverständlichkeit zugrunde liegt, dass das zu annotierende Material aus Sätzen bestehe – was für die Gesprächsdaten in FOLK aber nicht gilt.

Schließlich erfordern Gesprächskorpora auch vom Kernstück der korpuslinguistischen Analyse – der Korpus-Query, also der gezielten und systematischen Suche nach sprachlichen Formen im Korpus – weitreichende Anpassungen. Fast allen gängigen Recherchesystemen (wie dem Corpus Query Processor CQP, dem Corpus Search, Management and Analysis System COSMAS oder der Korpusanalyseplattform KorAP) liegt ein Modell zugrunde, das Korpora als eine Menge von Texten, ggf. mit diesen zugeordneten Metadaten, behandelt und die Texte selbst als „Stream of Tokens" (Menke et al. 2015) – also als lineare Abfolge von Wort- und Interpunktions-Tokens, ggf. mit Annotationen, die einzelne Tokens oder Token-Folgen referenzieren – betrachtet. In der Anwendung auf Gesprächsdaten greift ein solches Modell in mehrfacher Hinsicht zu kurz:

– Die Transkripte, auf denen eine Korpusrecherche im Falle von Gesprächskorpora ausgeführt wird, sind Sekundärdaten, die abschnittsweise den Primärdaten – also den Audio- oder Videoaufnahmen – zugeordnet sind. Bei der Recherche selbst kann dieser Umstand zunächst in den Hintergrund gerückt werden, bei der Präsentation des Rechercheergebnisses muss die Zuordnung aber nutzbar gemacht werden, indem ein Zugriff auf den betreffenden Abschnitt der Aufnahme ermöglicht wird.
– Nicht alle Metadaten beziehen sich auf den „Text" (d. h. das Transkript bzw. die zugrunde liegende Gesprächsaufnahme) als Ganzes. Soziobiographische Daten der Sprecher, die für viele Analysezwecke sehr wichtig sein

können, müssen jeweils nur den Beiträgen des betreffenden Sprechers – und damit nur ausgewählten „Text"-Teilen – zugeordnet werden. In einer Korpusrecherche muss diese Zuordnung nutzbar sein, z. B. indem eine Datenabfrage auf die Beiträge männlicher Sprecher aus dem norddeutschen Raum beschränkt wird.

- Neben vollwertigen Worttokens enthalten Transkripte auch andersartige Elemente, z. B. Beschreibungen nonverbalen Verhaltens, Pausen oder unvollständige Wörter (Abbrüche u. Ä.). Je nach Recherche-Interesse kann es sinnvoll sein, solche Elemente bei einer Query zu berücksichtigen oder unbeachtet zu lassen. Dies hat z. B. Auswirkungen auf die Berechnung von Token-Abständen in kontextsensitiven Suchen oder auf die Berechnung von Token-Frequenzen.
- Wort- und andere Tokens sind in Gesprächskorpora nicht durchgängig linear angeordnet – die Reihenfolge zweier Tokens aus überlappenden Redebestandteilen verschiedener Sprecher ist z. B. nicht immer eindeutig festgelegt, oder zwei solcher Tokens können identische Positionen haben. Ein echter „Stream of Tokens" kann daher immer nur lokal (für einzelne Sprecherbeiträge) oder für eine Teilmenge der Daten (alle Beiträge eines Sprechers) angenommen werden. Auch dies hat Konsequenzen z. B. für die Berechnung von Token-Abständen.
- Im selben Sinne kann der Begriff „Kontext" im Falle von Gesprächskorpora nicht auf die Tokens reduziert werden, die einem Recherchetreffer unmittelbar vorausgehen und folgen. Diese können zwar innerhalb einer Keyword-in-Context (KWIC)-Darstellung eines Rechercheergebnisses für einen kompakten Überblick genutzt werden. Zusätzlich muss aber die Möglichkeit gegeben sein, vorausgehende und folgende Beiträge des gleichen oder eines anderen Sprechers mit einzubeziehen, indem z. B. der gesamte zugehörige Transkriptausschnitt angezeigt werden kann.

Da sich insbesondere die Gesprächsanalyse vornehmlich für Strukturen und Mechanismen interaktiven Handelns interessiert, ist mit der „Textstruktur" (d. h. der in den Transkripten abgebildeten Struktur von Turn-Organisation, Sprecherwechseln etc.) schließlich auch eine weitere Dimension in der Korpusrecherche von Interesse, die theoretisch zwar auch für schriftsprachliche Texte von Belang ist, von den meisten gängigen Korpussystemen aber unberücksichtigt gelassen wird. Es geht hierbei um die Einschränkung einer Suche auf bestimmte strukturelle Positionen, beispielsweise „unmittelbar nach einem Sprecherwechsel", „innerhalb einer Überlappung" oder „in einem Beitrag mit weniger als drei Tokens".

In der Summe führen diese zusätzlichen Anforderungen wiederum dazu, dass korpuslinguistische Recherchesysteme für die Anwendung auf Gesprächs-

Abb. 9.2: Formulierung einer schrittweisen Suche auf FOLK in der DGD nach Vorkommen von „nein“, geäußert von männlichen Sprechern im unmittelbaren Kontext von „aber“ und am Beginn eines Sprecherbeitrags.

korpora umfassend angepasst oder eigens entwickelt werden müssen. Transkripte als einfache „Texte“ mit den Mechanismen der Systeme zu verarbeiten, die auf schriftsprachliche Texte ausgelegt sind, ist zwar möglich und wird auch praktiziert, z. B. bei der Integration des BNC in das Korpusportal der Brigham Young University (https://corpus.byu.edu [letzter Zugriff: 26. 9. 2017]). Ohne die Möglichkeiten eines Rückgriffs auf Audio- und Videoaufnahmen, einer Einschränkung von Suchen auf sprecherspezifische Metadaten oder der Berücksichtigung der Besonderheiten von Gesprächs-Tokens und deren Kontexteigenschaften bleibt aber ein Großteil des besonderen Potenzials von Gesprächskorpora ungenutzt. Ansätze, die diese Bedarfe adressieren, finden sich z. B. bei KonText, der Query Engine des *Czech National Corpus* (https://kontext.korpus.cz [letzter Zugriff: 26. 9. 2017]), im polnischen Portal SPOKES (http://spokes.clarin-pl.eu/ [letzter Zugriff: 26. 9. 2017]) oder in der französischen CLAPI-Plattform (http://clapi.ish-lyon.cnrs.fr [letzter Zugriff: 26. 9. 2017]). Für die Recherche auf FOLK in der DGD (http://dgd.ids-mannheim.de [letzter Zugriff: 26. 9. 2017]) wird ein schrittweise verfeinerbares Verfahren der Korpusrecherche angeboten, in dem gesprächsstrukturelle Constraints, Eigenschaften von Tokens und Tokens im Kontext sowie sprecherspezifische Metadaten einbezogen und miteinander kombiniert werden können (Abb. 9.2).

Bei der Präsentation der Ergebnisse als KWIC-Konkordanz gibt es dann die Möglichkeit, einzelne Treffer im Kontext des Transkriptausschnittes anzuzeigen und das zugrunde liegende Audio oder Video abzuspielen (Abb. 9.3).

Ergebnisse 1 bis 20 von 51 (51 / 0 aus-/abgewählt) Seite 1 von 3

	Ereignis	Sprecher	Treffer		Geschlecht
1	FOLK_00001	LB	nee	aber sie ham s verstanden denk isch	Männlich
2	FOLK_00007	JK	nee	aber man kann s ja kontrollieren ja un bevor ich	Männlich
3	FOLK_00011	VK	nein	aber weil ihr hier die ganze zeit so rum	Männlich
4	FOLK_00012	VK	nee	aber hier fängt ma an eins zwei drei zum beispiel	Männlich
5	FOLK_00021	SK	nee	aber dass die den direkt nach	Männlich
6	FOLK_00021	CH	nee	aber beim ha es vau bald	Männlich
7	FOLK_00021	MT	nee	aber	Männlich
8	FOLK_00021	SK	nö	aber des könnt ja sein	Männlich
9	FOLK_00021	CH	nö	aber ich könnt ja dann	Männlich

0587	CH	gib gib mir dann au ma (.) eine
0588	JZ	(.) bist du ferdig
0589		(0.56)
0590	CH	nö aber ich könnt ja (dann)
0591		(0.19)
0592	CH	nebenher schon ma anfangen
0593	XM1	((lacht))

	Ereignis	Sprecher	Treffer		Geschlecht
10	FOLK_00030	PB	nee	aber mit internetseiten meists manchmal	Männlich
11	FOLK_00039	NO	nee	aber is schon	Männlich
12	FOLK_00039	NO	nee	aber dass dass schon	Männlich
13	FOLK_00039	NO	nee	aber dann suchen wa uns en foto en foto suchen	Männlich
14	FOLK_00042	LK	nee	aber die	Männlich
15	FOLK_00042	LK	nein	aber ich nein nein nein nein nein so hab ich	Männlich
16	FOLK_00043	PB	nee	aber ich hab den dir hier mitgebracht	Männlich
17	FOLK_00043	PB	nee	aber ich mag diese nusspampe net so	Männlich
18	FOLK_00047	PB	nee	aber das ja ich will das ja auch f bisschen	Männlich
19	FOLK_00047	PB	nein	aber ku ma die sind alle reserviert	Männlich
20	FOLK_00066	PA	nee	aber ähm	Männlich

Ergebnisse 1 bis 20 von 51 (51 / 0 aus-/abgewählt) Seite 1 von 3

Abb. 9.3: KWIC-Konkordanz als Ergebnis zur Recherche aus Abbildung 9.2 mit eingeblendetem Transkriptausschnitt zu Treffer #9. Durch Doppelklick auf ein beliebiges Wort im Transkript oder Klick auf den „Play"-Button einer beliebigen KWIC-Zeile wird das zugehörige Audio abgespielt.

Zu den Herausforderungen für die Arbeit mit Gesprächskorpora gehört auch, solche und weitere Verfahren, die speziell den Interaktions-Aspekt der Sprache korpuslinguistisch zugreifbar machen, weiter zu entwickeln und zu verfeinern.

6 Schlussbemerkung

Gesprächskorpora sind in diesem Beitrag als ein besonderer Korpustyp definiert und beschrieben worden, dessen Potenzial für die Korpuslinguistik sich erst in jüngerer Zeit zu erschließen begonnen hat und für den sich methodische und technische Herausforderungen stellen, die über den korpuslinguistischen „Mainstream" hinausweisen. Wenn der „eklatante[n] Unterrepräsentation spontansprachlicher Daten in Korpora" (Mair in diesem Band) entgegengewirkt werden soll, können und sollten Gesprächskorpora dabei eine wichtige Rolle spielen. Die Korpuslinguistik (als linguistische Teildisziplin oder teildisziplinen-übergreifende methodische Ausrichtung) wäre dann gefordert, die Eigenheiten dieses „besonderen" Korpustyps in ihre Methodik einzubeziehen und ihre technologischen Lösungen so zu gestalten, dass Text-, Gesprächs- und ggf. weitere Korpustypen auf einer gemeinsamen Basis verarbeitet und analysiert werden können. Umgekehrt müssten angesichts

des großen Aufwandes, den die Erstellung von Gesprächskorpora mit sich bringt, Fragen der technischen Aufbereitung, der Standardisierung und des Teilens und Nachnutzens von Gesprächsdaten im Rahmen der gesprächsanalytischen Disziplinen verstärkte Aufmerksamkeit finden.

Literatur

Batinic, Dolores & Thomas Schmidt (2017): Reconstruction of separable particle verbs in a corpus of spoken German. Erscheint in: Proceedings der GSCL-Tagung 2017, Berlin.

Berens, Franz-Josef, Karl-Heinz Jäger, Gerd Schank & Johannes Schwitalla (1976): *Projekt Dialogstrukturen. Ein Arbeitsbericht* (Heutiges Deutsch I/12). München: Hueber.

Biber, Douglas (1993): Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4), 243–257.

Brinckmann, Caren (2009): Transcription bottleneck of speech corpus exploitation. In Verena Lyding (Hrsg.), *LULCL II 2008 – Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics. Bozen-Bolzano, 13th–14th November 2008*, 165–179. Bozen-Bolzano: EURAC.

Brünner, Gisela (1987/2005): *Kommunikation in institutionellen Lehr-Lern-Prozessen. Diskursanalytische Untersuchungen zu Instruktionen in der betrieblichen Ausbildung*. Tübingen: Narr; Neuauflage: Radolfzell: Verlag für Gesprächsforschung, 2005. http://www.verlag-gespraechsforschung.de/2005/bruenner.htm (letzter Zugriff: 26. 9. 2017).

Bührig, Kristin, Ortrun Kliche, Bernd Meyer & Birte Pawlack (2012): The corpus “Interpreting in Hospitals”: Possible approaches for research and communication training. In Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14), 305–314. Amsterdam: Benjamins.

Deppermann, Arnulf & Martin Hartung (2012): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch“ (FOLK) am Institut für Deutsche Sprache (Mannheim). In Ekkehard Felder, Marcus Müller & Friedemann Vogel (Hrsg.), *Korpuspragmatik*, 414–450. Berlin, Boston: de Gruyter.

Deppermann, Arnulf & Thomas Schmidt (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik – Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). *Mitteilungen des Deutschen Germanistenverbandes* 61 (1), 4–17.

Deppermann, Arnulf, Nadine Proske & Arne Zeschel (Hrsg.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr.

Dister, Anne, Françoise Gadet, Ralph Ludwig, Chantal Lyche, Lorenza Mondada, Stefan Pfänder, Anne Catherine Simon & Ingse Skattum (2008): Deux nouveaux corpus internationaux du français: CIEL-F (Corpus International et Écologique de la Langue Française) et CFA (Français contemporain en Afrique et dans l'Océan Indien). *Revue de Linguistique Romane* 285/286, 295–314.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, & Nii Martey (2000–2005). *Santa Barbara corpus of spoken American English, Parts 1–4*. Philadelphia, PA: Linguistic Data Consortium.

Ehlich, Konrad & Jochen Rehbein (1986): *Muster und Institution: Untersuchungen zur schulischen Kommunikation*. Tübingen: Narr.

Eshkol-Taravella Iris, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua & Isabelle Tellier (2012): Un grand corpus oral «disponible»: le corpus d'Orléans 1968–2012. *Ressources Linguistiques Libres, TAL* 52 (3), 17–46.

Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (2012): The GeWiss corpus: Comparing spoken academic German, English and Polish. In Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14), 319–338. Amsterdam: Benjamins.

Fandrych, Christian, Elena Frick, Hanna Hedeland, Anna Iliash, Daniel Jettka, Cordula Meißner, Thomas Schmidt, Franziska Wallner, Kathrin Weigert & Swantje Westpfahl (2016): User, who art thou? User profiling for oral corpus platforms. In Nicoletta Calzolari et al. (Hrsg,), *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, 280–287. Paris: European Language Resources Association (ELRA).

Glas, Reinhold & Konrad Ehlich (2000): Deutsche Transkripte 1950 bis 1995. Ein Repertorium (Arbeiten zur Mehrsprachigkeit, 63). Hamburg: Institut für Germanistik I/Arbeitsbereich Deutsch als Fremdsprache, Arbeitsstelle Mehrsprachigkeit/Research Center for Multilingualism, Universität Hamburg.

Groupe ICOR (Michel Bert, Sylvie Bruxelles, Carole Etienne, Lorenza Mondada, Véronique Traverso) (2010): Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). *Pratiques – Interactions et corpus oraux* 147–148, 17–34.

Haugh, Michael & Wei-Lin M. Chang (2013): Collaborative creation of spoken language corpora. In Tim Greer, Donna Tatsuki & Carsten Roever (Hrsg.), *Pragmatics and Language Learning, Volume 13*, 133–159. Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.

Herkenrath, Annette & Jochen Rehbein (2012): Pragmatic corpus analysis, exemplified by Turkish-German bilingual and monolingual data. In: Thomas Schmidt & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies in Multilingualism 14) 123–152. Amsterdam: Benjamins.

Hohenstein, Christiane (2006): *Erklärendes Handeln im Wissenschaftlichen Vortrag. Ein Vergleich des Deutschen mit dem Japanischen* (Studien Deutsch 36). München: iudicium.

Kaiser, Julia (2016): Reformulierungsindikatoren im gesprochenen Deutsch: Die Benutzung der Ressourcen DGD und FOLK für gesprächsanalytische Zwecke. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 17, 196–230.

Kallmeyer, Werner (Hrsg.) (1994): *Exemplarische Analysen des Sprachverhaltens in Mannheim. Kommunikation in der Stadt* (Schriften des Instituts für deutsche Sprache 4.1). Berlin, New York: de Gruyter.

Keim, Inken (2008): *Die „türkischen Powergirls" – Lebenswelt und kommunikativer Stil einer Migrantinnengruppe in Mannheim* (Studien zur deutschen Sprache 39). 2. korrig. Aufl. Tübingen: Narr.

Koch, Peter & Wulf Oesterreicher (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15–43.

Kupietz, Marc & Thomas Schmidt (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In Ludwig M. Eichinger (Hrsg.):

Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven (Jahrbuch des Instituts für Deutsche Sprache 2014), 297–322. Berlin, Boston: de Gruyter.

Labov, William (1972): *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania.

Lalouschek, Johanna & Florian Menz (2002): Empirische Datenerhebung und Authentizität von Gesprächen – am Beispiel medizinischer Kommunikation. In Gisela Brünner, Reinhard Fiehler & Walther Kindt (Hrsg.), *Angewandte Diskursforschung*, Band 1: *Grundlagen und Beispielanalysen*, 46–68. Radolfzell: Verlag für Gesprächsforschung.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery (2017): The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22 (3), 319–344.

Luckmann, Thomas (1986): Grundformen der gesellschaftlichen Vermittlung des Wissens: Kommunikative Gattungen. Zeitschrift für Soziologie 27, 191–211.

Menke, Peter, Farina Freigang, Thomas Kronenberg, Sören Klett & Kirsten Bergmann (2015): First steps towards a tool chain for automatic processing of multimodal corpora. *Journal of Multimodal Communication Studies* 2, 30–43. http://jmcs.home.amu.edu.pl/wp-content/uploads/2015/09/Menke_et_al_2014_JMCS.pdf.

Meliss, Meike & Christine Möhrs (2017): Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. *Sprachreport* 4/2017, 42–53. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-68549

Moore, Robert J. (2015): Automated transcription and conversation analysis. Research on Language and Social Interaction 48 (3), 253–270.

Nothdurft, Werner, Ulrich Reitemeier & Peter Schröder (1994): *Beratungsgespräche. Analyse asymmetrischer Dialoge* (Forschungsberichte des Instituts für deutsche Sprache 61) Tübingen: Narr.

Ochs, Elinor (1979): Transcription as theory. In Elinor Ochs & Bambi Schieffelin (Hrsg.) (1979), *Developmental pragmatics*, 43–72. New York u. a.: Academic Press.

Oostdijk, Nelleke (2002): The design of the Spoken Dutch Corpus. In Pam Peters, Peter Collins & Adam Smith (Hrsg.), *New frontiers of corpus research*, 105–112. Amsterdam: Rodopi.

Rehbein, Jochen & Petra Löning (1993): *Arzt-Patienten-Kommunikation: Analysen zu interdisziplinären Problemen des medizinischen Diskurses*. Amsterdam: Benjamins.

Rehbein, Jochen, Jutta Fienemann, Sören Ohlhus & Christine Oldörp (2001): Nonverbale Kommunikation im Videotranskript. Zu nonverbalen Aspekten höflichen Handelns in interkulturellen Konstellationen und ihre Darstellung in computergestützten Videotranskriptionen. In Dieter Möhn, Dieter Roß & Marita Tjarks-Sobhani (Hrsg.), *Mediensprache und Medienlinguistik. Festschrift für Jörg Hennig*, 167–198. Frankfurt am Main: Peter Lang.

Schmidt, Thomas (2005a): *Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit text-technologischen Mitteln*. Frankfurt am Main: Peter Lang.

Schmidt, Thomas (2005b): Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 6, 103–126.

Schmidt, Thomas (2014a): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. Gesprächsforschung – *Online-Zeitschrift zur verbalen Interaktion* 15, 196–233.

Schmidt, Thomas (2014b): The database for spoken German – DGD2. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Ninth International Conference on Language Resources*

and Evaluation (LREC'14), Reykjavik, Iceland, 1451–1457. Reykiavik: European Language Resources Association (ELRA).

Schmidt, Thomas (2016a): Good practices in the compilation of FOLK (Research and Teaching Corpus of Spoken German). *International Journal of Corpus Linguistics* 21 (3), 396–418.

Schmidt, Thomas (2016b): Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics* 31 (1), 127–154.

Schmidt, Thomas & Kai Wörner (Hrsg.) (2012): *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies on Multilingualism 14). Amsterdam: Benjamins.

Schröder, Peter (1975): Die Untersuchung gesprochener Sprache im Projekt ‚Grundstrukturen der deutschen Sprache' – Planungen, Probleme, Durchführung. In Ulrich Engel & Irmtraud Vogel (Hrsg.), *Gesprochene Sprache: Bericht der Forschungsstelle Freiburg* (Forschungsberichte des Instituts für Deutsche Sprache 7), 5–46. 2. Aufl. Tübingen: Narr.

Schröder, Peter (Hrsg.) (1997): *Schlichtung*, Band 3: *Schlichtungsgespräche. Ein Textband mit einer exemplarischen Analyse* (Schriften des Instituts für Deutsche Sprache 5.3). Berlin, New York: de Gruyter.

Schütte, Wilfried (1996): Boulevardisierung von Information: Streitgespräche und Streitkultur im Fernsehen. In Bernd Ulrich Biere & Rudolf Hoberg, (Hrsg.), *Mündlichkeit und Schriftlichkeit im Fernsehen* (Studien zur deutschen Sprache 5), 101–134. Tübingen: Narr.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek & Marko Stabej (2013): Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation* 47 (4), 1031–1048, doi: 10.1007/s10579-013-9216-5.

Wagener, Peter & Karl-Heinz Bausch (Hrsg.) (1997): *Tonaufnahmen des gesprochenen Deutsch. Dokumentation der Bestände von sprachwissenschaftlichen Forschungsprojekten und Archiven*. Berlin, New York: de Gruyter.

Westpfahl, Swantje & Thomas Schmidt (2016): FOLK-Gold – A GOLD standard for part-of-speech-tagging of spoken German. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, 1493–1499. Paris: European Language Resources Association (ELRA).

Wiese, Heike, Ulrike Freywald, Sören Schalowski & Katharina Mayr (2012): Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache* 2, 97–123.

Wolfgang Imo und Beate Weidner

10 Mündliche Korpora im DaF- und DaZ-Unterricht

Abstract: In dem vorliegenden Beitrag geht es darum, die Möglichkeiten und Vorteile des Einsatzes von Korpora mit mündlichen Sprachdaten im DaF- und DaZ-Unterricht darzustellen. Dabei wird auf die Frage eingegangen, inwieweit ‚authentische' gesprochene Sprache ihren Platz im Fremdsprachenunterricht hat und welche Korpora für Unterrichtszwecke zur Verfügung stehen, um schließlich etwas ausführlicher auf die *Plattform Gesprochenes Deutsch* einzugehen, die speziell an den Bedürfnissen von Lehrenden und Lernenden in den Bereichen DaF und DaZ ausgerichtet ist.

Keywords: Alltagssprache, DaF, DaZ, gesprochene Sprache, interaktionale Sprache, Sprachkorpora

1 Einleitung

Selbstverständlich ist die Berücksichtigung mündlicher Alltagsinteraktion im DaF- und DaZ-Unterricht nicht erst seit dem Gemeinsamen Europäischen Referenzrahmen (2001) ein in der Fremdsprachendidaktik diskutiertes Thema. Mündliche Kommunikation gehört zum Lernen einer Sprache genauso dazu wie schriftliche Kommunikation, wobei gerade die technologische Entwicklung (von Lernschallplatten über Sprachlabore bis zu Online-Ausspracheübungen, um nur einige zu nennen) stets Vorreiter für den Ausbau der Thematisierung und Lehre mündlicher Strukturen im DaF- und DaZ-Unterricht war. So stellen Hirschfeld, Rösler & Schramm (2016) fest,

> dass das Pendel in der Methodengeschichte der Fremdsprachendidaktik immer wieder zwischen Schriftlichkeit und Mündlichkeit hin- und hergeschwungen ist, dass sich die Medien im Lauf der Jahrzehnte mit immer höherer Geschwindigkeit vom Tonband über die Audiokassette hin zu digitalen Dateien, Podcasts und Skype-Anrufen verändert haben

Wolfgang Imo, Universität Hamburg, Institut für Germanistik, Überseering 35, D-22297 Hamburg, E-Mail: Wolfgang.Imo@uni-hamburg.de
Beate Weidner, Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6–8, D-45127 Essen, E-Mail: Beate.Weidner@uni-due.de

https://doi.org/10.1515/9783110538649-011

> und dass der fremdsprachendidaktische Blick auf die Mündlichkeit in manchen Zeiten stärker das Sprechen und in anderen Zeiten stärker das Hören in den Mittelpunkt gestellt hat. (Hirschfeld, Rösler & Schram 2016: 132–133)

Mit dem zunehmenden Aufbau und der Bereitstellung von Gesprächsdatenbanken in Verbindung mit deren einfachem Zugang über das Internet ist eine neue technische Stufe eingetreten, die es ermöglicht, nun auch den großen Bereich authentischer Gespräche – und nicht nur speziell für den Unterricht erstellte Interaktionen – im Unterricht einzusetzen. Im englischsprachigen Raum werden bereits seit längerer Zeit die Möglichkeiten des Einsatzes von Sprachkorpora im Sprachunterricht diskutiert (vgl. den Sammelband *How to use corpora in language teaching* von Sinclair [2004]). Für das Deutsche liegen aktuelle Arbeiten u. a. von Fandrych & Tschirner (2007), Siepmann (2009), Lüdeling & Walter (2010) und Bartz & Radtke (2014) mit einem dezidierten Blick auf den Einsatz von Korpora im Unterricht vor.

In dem vorliegenden Beitrag soll es allerdings nicht darum gehen, Sinn und Zweck dezidiert korpuslinguistischer Methoden für den DaF- und DaZ-Unterricht zu diskutieren. Die Frage ist vielmehr, welche Korpora gesprochener Sprache speziell für diese Bereiche überhaupt existieren, wie einfach sie zugänglich sind und wie sie jenseits von Fragestellungen der Korpuslinguistik im engeren Sinne als Unterrichtsressourcen verwendet werden können. Das führt dazu, dass zunächst geklärt werden muss, welchen Zweck die Thematisierung von authentischer, informeller gesprochener Sprache im DaF- und DaZ-Unterricht überhaupt hat und wo die Vor- und Nachteile davon liegen (Abschnitt 2), bevor dann ein kurzer kritischer Überblick über bestehende Ressourcen gegeben wird (Abschnitt 3). In Abschnitt 4 wird detailliert eine im Aufbau befindliche, speziell an DaF- und DaZ-Bedürfnisse adressierte Gesprächsdatenbank vorgestellt. Abschließend werden anhand von Beispielen aus der Datenbank ausgewählte Phänomene interaktionaler Sprache mit Anwendungsbezug für den DaF- und DaZ-Unterricht analysiert.

2 Wieso authentische gesprochene Sprache im DaF/DaZ-Unterricht?

Das Stichwort „Authentizität" fällt im Bereich des Fremdsprachenunterrichts immer wieder, denn Authentizität steht als „Begriff für das Gebot, von Muttersprachlern verfasste oder gesprochene Texte zu verwenden anstatt solcher, die im Fremdsprachenunterricht, meist von Nicht-Muttersprachlern, eigens für den Fremdsprachenunterricht hergestellt oder bearbeitet werden" (Edelhoff 1985: 7;

vgl. auch Civegna 2005). Authentische Texte gelten dabei als das Gegenstück zu den klassischen, dem Niveau der Lernenden angepassten und stets auch bestimmte Zwecke der Sprachvermittlung (z. B. das Üben von Frageformaten, des Perfekts, des Passivs etc.) verfolgenden Lehrbuchtexten. Der Vorteil authentischer Texte liegt darin, dass sie eine höhere Realitätstreue aufweisen, d. h. sie bereiten auf den tatsächlichen mündlichen oder schriftlichen Kontakt mit der Zielsprache vor und können daher auch zu einer höheren Lernmotivation beitragen. Unter authentischen Sprachdaten können ganz allgemein diejenigen gefasst werden, „die nicht eigens für didaktische Zwecke erstellt wurden“ (Lüger 2009: 15). Authentische *Gespräche* im Besonderen sind solche, die „nicht extra zum Zweck der Untersuchung geführt oder inszeniert wurden; es werden also natürliche Gespräche aus dem Alltags- und Berufsleben untersucht.“ (Becker-Mrotzek & Brünner 2006: 3) Für den Bereich der gesprochenen Sprache umfasst diese Definition somit nicht nur interaktionale und umgangssprachliche Gesprächstypen wie Familiengespräche, Gespräche unter Freunden und Bekannten o. Ä., sondern auch medial vermittelte Gesprächstypen wie Nachrichten im Radio oder Fernsehen oder Talk-Sendungen. Umstritten ist, inwieweit auch Gespräche beispielsweise aus Spielfilmen noch als authentisch bezeichnet werden können: Einerseits sind sie nicht für didaktische Zwecke erstellt, aber andererseits sind sie insofern nicht „natürlich“, als sie mehr oder weniger stark „gescriptet“ sind (vgl. beispielsweise Reershemius & Ziegler [2015] zur Diskussion der Darstellung von ethnolektalem Deutsch in dem Film *Fack ju Göhte*).

Stellt man nun die Frage, in welchen Bereichen im Fremdsprachenunterricht ‚authentische‘ Sprachdaten zum Einsatz kommen können und wie dieser bewertet wird, so ergibt sich (vor allem im DaF-Bereich) ein gemischtes Bild: Der Einsatz authentischer *geschriebener* Sprache im Fremdsprachenunterricht hat eine lange Tradition und liegt auch auf der Hand: Einerseits werden Zeitungstexte und Artikel aus Zeitschriften eingesetzt, da mit ihnen zugleich auch Aspekte der Landeskunde (politisches System, Gesellschaftsstruktur etc.) thematisiert werden können, und andererseits werden literarische Texte (Gedichte, Auszüge aus Romanen oder Erzählungen etc.) verwendet, um damit neben der Sprache zugleich auch literarische Kultur zu vermitteln. Gegenüber dem Einsatz authentischer gesprochener Sprache – und dabei vor allem interaktionaler, informeller Sprache – herrschen dagegen Vorbehalte: Für die Thematisierung von Umgangssprache sei weder Zeit (diese werde für die Bereiche der Standardvermittlung benötigt), noch sei es wünschenswert, den Lernenden ‚fehlerhaftes‘ Deutsch zu vermitteln. Je näher an der Standardschriftsprache die Gesprächsdaten sind, desto eher ist die Bereitschaft zu erkennen, diese Daten auch im Unterricht zu verwenden (vgl. beispielsweise den Vorschlag von Mac [2011], Fernsehnachrichten als Lehr- und Arbeitsmaterial einzusetzen).

Bei „alltäglicher“ Sprache herrscht dagegen die Sichtweise vor, dass ein Zuviel an Authentizität eher schaden könne. Günthner, Wegner & Weidner (2013) zeigen dies beispielhaft an einer Analyse eines Lehrbuchdialogs aus einem DaF-Lehrwerk (Stufe B2), der durch den Einsatz von ungebräuchlichen Konjunktivformen und gestelzt wirkenden Formulierungen bei gleichzeitiger Verwendung von einigen – relativ unmotiviert gesetzten – umgangssprachlichen Mustern geradezu absurd wirkt. Lernenden ist damit wenig geholfen, da so der Eindruck entsteht, dies sei tatsächlich die Art und Weise, wie man im Alltag in Deutschland kommuniziert (vgl. auch Bachmann-Stein [2013: 41 f.] zu Problemen fingierter Dialoge).

Es ist natürlich klar, dass es nicht sinnvoll ist, auf allen Lernstufen mit authentischen Sprachdaten zu arbeiten. Lerntexte und Lerndialoge sind ein bewährtes Mittel, das der Vermittlung von „transitorischen Normen“ (Feilke 2012: 155) dient, d. h. von Normen, die im Sinne eines „Scaffolding“ Sprachkompetenz schrittweise aufbauen und die bei zunehmender Sprachbeherrschung entsprechend situationsgebunden auch aufgegeben oder verändert werden können. Umgekehrt heißt dies aber auch, dass authentische gesprochene Sprache – um diesen für den vorliegenden Beitrag relevanten Bereich zu fokussieren – mit zunehmender Sprachkompetenz lernrelevant werden muss: Nur durch den Kontakt mit der Art und Weise des Sprechens, wie sie in Deutschland zum unmarkierten Umgang gehört, können die Lernenden das Einordnen der transitorischen Normen als ebensolche erlernen. Damit zusammenhängend erwerben sie auch die notwendige ‚Spielkompetenz‘, die darin besteht, situationsspezifisch entscheiden zu können, wie und in welchem Umfang sich an Normen orientiert werden muss. Im DaZ-Unterricht ist dies sogar noch dringender notwendig als im DaF-Unterricht, da der Kontakt mit Alltagssprache von Beginn an in hohem Maße vorhanden ist. Im DaF-Unterricht dagegen spielt er je nach Ausgangsland und Lernzielen eine eher untergeordnete Rolle (z. B. wenn der Aufenthalt erst zum Ende des Bachelor-Studiums erfolgt bzw. Deutsch zum Zweck des Übersetzens von standardnahen schriftlichen Texten erworben wird).

Der Gemeinsame Europäische Referenzrahmen für Sprachen (2001) lässt sich zudem mit seinen Forderungen nach Handlungsorientierung und kommunikativen Kompetenzen explizit als Aufforderung interpretieren, nicht nur didaktisierte Dialoge, sondern auch authentische Gespräche im Unterricht einzusetzen (vgl. Imo 2011, 2012, 2013a, 2013b, 2015, 2016, i. V.; Weidner 2018). Ganz nebenbei besteht der Vorteil solcher Daten auch darin, den „Praxisschock“ zu verringern, der im Gegensatz zum DaZ-Bereich im DaF-Bereich unvermeidbar ist, wenn die Lernenden erstmals mit gesprochenem Alltagsdeutsch in Berührung kommen und erkennen, dass sie sowohl Probleme mit der Geschwindigkeit, der Ausspracherealisierung dem Verstehen von um-

gangssprachlichen oder dialektalen Ausdrücken etc., haben als auch Probleme bei der Produktion von Äußerungen.[1] Im DaF-Bereich gewinnt diese Einsicht zunehmend an Raum: Nicht nur gestützt durch die Aufnahme – und damit Aufwertung – von Strukturen gesprochener Sprache in die Duden Grammatik (seit der 7. Auflage 2006), sondern auch durch zahlreiche Arbeiten der letzten Jahre, die den Einsatz authentischer Gesprächsdaten im DaF-Bereich thematisieren (Günthner 2000, 2011; Handwerker, Bäuerle & Sieberg 2016; Hennig 2002; Imo 2009, 2011, 2012, 2013c, i. V.; Imo & Moraldo 2015; Moraldo 2011; Moraldo & Missaglia 2013; Pieklarz 2010; Pieklarz-Thien 2015; Reeg, Gallo & Moraldo 2012; Sieberg 2013 und Weidner 2012), ist inzwischen der Boden bereitet für die Verwendung von Datenbanken mit authentischen gesprochensprachlichen Daten im Fremdsprachenunterricht.[2] Auch im Bereich DaZ wird in den curricularen Vorgaben die Vermittlung von alltagstauglichen Kommunikationskompetenzen zum obersten Ziel des Unterrichts erhoben. So nennt etwa der Lehrplan für Vorbereitungsgruppen „Deutsch als Zweitsprache“ des Sächsischen Staatsministeriums für Kultus (2009) als Lernziel, „die alltagssprachlichen Ausdrucksweisen verfügbar zu machen“. Auch der Bildungsplan für Stadtteilschulen der Freien und Hansestadt Hamburg (2011) gibt für Deutsch als Zweitsprache in Vorbereitungsklassen die Leitlinie vor, die „sprachliche Handlungsfähigkeit der Schülerinnen und Schüler in der Alltagskommunikation“ zu fördern.

3 Korpora für DaF und DaZ

Im Folgenden soll ein kurzer Überblick über einige Korpora gegeben werden, die für den DaF- und DaZ-Unterricht eingesetzt werden können und die authentisches, interaktionales gesprochenes Deutsch enthalten. Der Fokus liegt also nicht auf Lernerkorpora wie beispielsweise dem *Hamburg Modern Times Corpus* (http://hdl.handle.net/11022/0000-0000-6973-9 [letzter Zugriff: 27. 1. 2017]) oder dem *Hamburg Map Task Corpus* (http://hdl.handle.net/11022/0000-0000-6330-A [letzter Zugriff: 27. 1. 2017]), d. h. auf Korpora, die Aufnahmen

1 Wittig (2015: 155) bringt dies auf den Punkt: „Den Studierenden muss zunächst vermittelt werden, dass kommunikative Kompetenz und die Fähigkeit, grammatisch korrekte Sätze zu bilden, nicht miteinander gleichzusetzen sind.“

2 Zudem entstehen parallel auch konzeptionell ausgearbeitete Lernermodelle wie z. B. Handwerker & Madleners (2009) „Chunks für DaF“, das durch die Fokussierung auf das Erlernen von situationseingebetteten Sequenzen, also größeren sprachlichen Einheiten, die auf die Erfüllung routinierter kommunikativer Aufgaben zugeschnitten sind, optimal für den Einsatz mit Gesprächsdaten geeignet ist.

von Lernenden des Deutschen als Fremd- oder Zweitsprache enthalten (eine Diskussion von Lernerkorpora und einschlägigen auf Analysen solcher Korpora beruhenden Arbeiten findet sich in Lüdeling & Walter 2010). Stattdessen wird auf solche Korpora fokussiert, die muttersprachliche Sprachdaten zur Verfügung stellen (ein aktueller Überblick über generell verfügbare deutschsprachige Korpora findet sich in Lemnitzer & Zinsmeister 2015: 147–148).

Eine besondere Klasse von Korpora sind die *phonetischen Lernerkorpora*. Diese sind meist so aufgebaut, dass sowohl Aufnahmen von Muttersprachlern als auch von Fremdsprachenlernern enthalten sind. Entsprechend werden diese Korpora zu großen Teilen für die Analyse von Besonderheiten des Spracherwerbs auf der phonetischen und prosodischen Ebene eingesetzt, seltener sind auch Anwendungen zum konkreten Einsatz im Fremdsprachenunterricht vorgesehen, wie z. B. im Kontext eines „induktiv-entdeckenden Lernens", bei dem Lernende Strukturen aus den Korpusdaten heraus selbst erarbeiten. Generell kommt diesen Korpora eher eine „unterstützende Rolle für die Unterrichtsvorbereitung und Materialienentwicklung" (Gut 2007: 2) zu. Beispiele für solche Korpora sind u. a. das *LeaP-Korpus*, das im Kontext des Projekts „Learning Prosody in Foreign Language" in Bielefeld aufgebaut wurde. Es enthält insgesamt 12 Stunden Audiomaterial, von denen die Mehrzahl Aufnahmen deutscher und englischer Fremdsprachenlernender sind und eine kleine Anzahl jeweils muttersprachliche Aufnahmen. Das Korpus eignet sich vor allem zum Phonetiktraining im Fremdsprachenunterricht (vgl. ausführlich Gut 2007, 2009). Das *Kiel Corpus* (www.ipds.uni-kiel.de/forschung/kielcorpus.de.html [letzter Zugriff: 27. 1. 2017]) stellt kostenpflichtige CDs mit Aufnahmen deutscher Muttersprachler mit Lese- und Spontansprache zur Verfügung, die auch im Unterricht zum Hör- und Aussprachetraining eingesetzt werden können. Das *Strange Corpus* des Bayerischen Archivs für Sprachsignale (BAS) (www.phonetik.uni-muenchen.de/Bas/BasSC10deu.html [letzter Zugriff: 27. 1. 2017]) verfolgt eher den Zweck von phonetischen Analysen und ist nicht für den Unterrichtseinsatz ausgelegt. Ein aktuelles phonetisches Lernerkorpus (das *IFCASL-Korpus; Individualized Feedback in Computer-Assisted Spoken Language Learning*), das „ab 2017 der wissenschaftlichen Öffentlichkeit zugänglich gemacht wird", wird von Trouvain & Zimmerer (2016) vorgestellt. Dieses Korpus enthält jeweils Lerner- und Muttersprachdaten und ist ebenfalls primär für analytische Zwecke erstellt worden. Der Einsatz der phonetischen Korpora ist auf Grund des starken Fokus auf Lernerdaten nicht für die Thematisierung von authentischem Deutsch geeignet (die muttersprachlichen Anteile sind meist nur gering) und bestenfalls für spezielle Teilfragen (Aussprachetraining, Prosodietraining) sinnvoll im Unterricht einsetzbar.

Das *Berlin Map Task Corpus (BeMaTaC)* (http://u.hu-berlin.de/bematac [letzter Zugriff: 27. 1. 2017]) ist ein experimentell angelegtes Korpus, das jeweils

aus zwei strukturell gleich aufgebauten Teilkorpora mit gleichen Aufgabensets besteht; einmal jeweils mit Muttersprachlern und einmal mit Lernenden von Deutsch als Fremdsprache. Dieses Korpus enthält zwar in einem Teilkorpus auch authentisches gesprochenes Deutsch, es ist aber auf Grund der künstlichen Experimentsituation einer Map Task problematisch. Das Korpus ist frei zugänglich, allerdings aber gerade auch wegen des Fokus auf komplexe korpusanalytische Aufbereitung eher für universitäre Zwecke und nicht allgemeinen für den DaF- und DaZ-Unterricht ausgerichtet.

Mit dem *GeWiss-Korpus* (https://gewiss.uni-leipzig.de [letzter Zugriff: 27.1. 2017]), das gesprochene Wissenschaftssprache enthält, liegt eine für den DaF- und DaZ-Unterricht sehr gut nutzbare Ressource vor. Auch wenn das Korpus von der Zielsetzung her eher als ein wissenschaftliches Vergleichskorpus von unterschiedlichen Wissenschaftssprachen geplant ist, können die deutschsprachigen Aufnahmen sehr gut im Unterricht eingesetzt werden. Allerdings ist der Anwendungsbereich insofern eingeschränkt, als die Daten vor allem für akademisch orientierte Lernende von Interesse sein werden, da die „Vermittlung des Deutschen als Wissenschaftssprache" (Fandrych, Meißner & Slavcheva 2014) durch die thematische Orientierung zwangsläufig im Mittelpunkt stehen muss. Auf Grund des vergleichsweise einfachen Zugangs (über eine Anmeldung auf der Homepage) und der leichten Bedienbarkeit ist ansonsten der Einsatz im Unterricht zu empfehlen.

Eine Korpussammlung, die gelegentlich als Ressource für den Unterricht – sowohl für den muttersprachlichen Unterricht, wie bei Bartz (2015) oder Bartz & Radtke (2014), als auch für den Fremdsprachenunterricht, wie bei Wallner (2013) – herangezogen wird, ist die des *Digitalen Wörterbuchs der Deutschen Sprache* (DWDS). Der Großteil dieser Korpora besteht allerdings aus Textsorten aus den Bereichen der Belletristik, der Gebrauchsliteratur, der Wissenschaft und der Zeitungssprache. Gesprochene Sprache ist prozentual deutlich unterrepräsentiert (vgl. die Tabellen auf der DWDS-Seite mit der Korpusbeschreibung unter www.dwds.de/r#group-Referenzkorpora [letzter Zugriff: 27.1. 2017]), worauf auch explizit hingewiesen wird: „Für die Textsorte ‚Gesprochene Sprache' konnte keine vollständige zeitliche Ausgewogenheit erreicht werden." Das Teilkorpus „Gesprochene Sprache" enthält „Transkripte aus dem gesamten 20. Jahrhundert", genannt werden u. a. „Reden, Rundfunkansprachen, Auszüge aus österreichischen Parlamentsprotokollen, Auszüge aus ca. 250 Spiegel-Interviews, Auszüge aus dem Literarischen Quartett von 1988 bis 2001, Auszüge aus dem Projekt Emigrantendeutsch in Israel sowie Auszüge aus Bundestagsprotokollen von 1998 bis 1999" (www.dwds.de/r#group-Referenzkorpora [letzter Zugriff: 27.1. 2017]). Die Tatsache, dass von „Textsorten" gesprochen wird, ist allerdings bezeichnend: In der Tat handelt es sich lediglich um Tran-

skripte ohne Audiodateien. Dies ist ein Mangel, der das Teilkorpus im DaF-/DaZ-Unterricht kaum sinnvoll einsetzbar macht, da es am Ziel der Thematisierung gesprochener Sprache vorbeigeht, sich beim Lehren und Lernen auf Transkripte ohne Audiodateien zu stützen, wie beispielsweise bei Bachmann-Stein (2013), wo Daten aus Transkriptbänden entnommen werden. Bei dem Einsatz von Transkripten im Unterricht ist stets darauf zu achten, dass, wie bei Liedtke (2013) beschrieben, der „Arbeit mit Transkripten" *immer* auch eine Arbeit mit den zugehörigen Audiodateien zur Seite gestellt wird.

Eine der bekanntesten und größten Datenbanken ist die *Datenbank Gesprochenes Deutsch* (DGD) (http://dgd.ids-mannheim.de/, [letzter Zugriff: 27.1.2017]) des Instituts für Deutsche Sprache (IDS), die seit 2012 ausgewählte Auszüge aus den im Laufe unterschiedlicher Forschungsprojekte gesammelten und archivierten Daten für Forschung und Lehre bereitstellt. Die Datenbank ist nach kostenfreier Registrierung zugänglich. Sie stellt umfangreiche Recherchemöglichkeiten sowohl in den Transkripten als auch in den Metadaten zur Verfügung und ermöglicht auch den Download von Einzelbelegen. Die meisten der Daten der DGD sind allerdings für Zwecke des Fremdsprachenunterrichts schon auf Grund ihres Alters (Pfeffer-Korpus der deutschen Umgangssprache, Zwirner-Korpus der deutschen Mundarten) nicht geeignet. Dies gilt jedoch nicht für das seit einiger Zeit aufgebaute und ständig weiterwachsende Teilkorpus *FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch)*. Dieses ist nicht nur aktuell, es strebt zudem eine Ausgewogenheit von Kommunikationssituationen an (Gespräche beispielsweise aus dem Arbeits-, Freizeit- und Bildungskontext), es werden außer Audiodaten auch Videoaufzeichnungen authentischer Gespräche bereitgestellt, und schließlich können aus dem FOLK-Korpus nicht nur einzelne Belegstellen, sondern komplette Datensätze, bestehend aus Audioaufnahme und Transkript, heruntergeladen werden, was die Weiterverarbeitung im Rahmen der Erstellung von Lehreinheiten erleichtert. Allerdings handelt es sich auch bei FOLK nicht um eine dezidierte Lehrdatenbank für den DaF- oder DaZ-Unterricht. Die umfangreichen Annotations- und Rechercheoptionen sowie die Tatsache, dass lediglich eine sehr grobe Minimaltranskription bereitgestellt wird, die im Rahmen eigener Forschungsarbeit dann entsprechend angepasst werden muss, zeigen, dass die Hauptadressaten im universitären Forschungsbereich liegen. Wer als Lehrender über gewisse gesprächsanalytische Vorkenntnisse verfügt, kann aber das FOLK durchaus gewinnbringend für Lehrzwecke einsetzen, wie beispielsweise Sieberg (2016) mit einem fremdsprachendidaktischen Vorschlag der Vermittlung von Responsiven und „Reaktiven" zeigt.

Als letztes – und die Überleitung zu Abschnitt 4 einleitend – kann die *Datenbank Gesprochenes Deutsch für die Auslandsgermanistik* (http://audiolabor.

uni-muenster.de/daf/ [letzter Zugriff: 27. 1. 2017]) genannt werden, die mit Unterstützung des DAAD im Jahr 2010 aufgebaut wurde. Diese Datenbank wurde explizit auf die Bedürfnisse von Lehrenden und Lernenden des Deutschen als Fremdsprache zugeschnitten und fokussiert nicht auf phonetische Aspekte, sondern auf die Strukturen gesprochener Sprache generell, wobei eine möglichst unkomplizierte Bedienung der Datenbank im Mittelpunkt stand. In der Datenbank können Gesprächsdaten im MP3-Format zusammen mit den dazugehörigen Transkripten im .doc- sowie .pdf-Format heruntergeladen werden. Darüber hinaus werden Informationsmaterialien (Fachartikel, die sich auf die Datenbank beziehen, ebenso wie Lehreinheiten) bereitgestellt. Eine Einschränkung ist allerdings die vom Förderer gesetzte Vorgabe, dass die Datenbank ausschließlich der Auslandsgermanistik zur Verfügung gestellt wird. Alle Gesprächsdaten sind ‚authentisch' in dem in Abschnitt 2 beschriebenen Sinn: Sie wurden nicht inszeniert und sie stammen aus keinem experimentellen Setting. Die Gesprächsausschnitte sind jeweils nur wenige Minuten lang, was sie für einen Einsatz im Unterricht prädestiniert. Zudem enthält jedes der Dokumente eine kurze Erläuterung der Transkriptionskonventionen sowie eine Situationsbeschreibung zu Beginn und eine Liste mit potenziell für Lernende problematischen Ausdrücken, die in den Gesprächen vorkommen. Damit sind die Gespräche für den unkomplizierten Einsatz im Unterricht vorbereitet. Die Ausschnitte sind thematisch und strukturell geordnet. Sie sind verschlagwortet und umfassen u. a. Themenbereiche wie „Sprechen über Fußball", „Urlaubsplanung", „Rezepte austauschen/Sprechen über Lebensmittel", „Thema Studium", „Gespräche im Friseursalon", „Lästergeschichten", „Erlebnisgeschichten", „Verabredungen treffen am Telefon", „Diskutieren", „Erklären", „Sprechstundengespräche an der Hochschule", „Verkaufsgespräche und Beratungsgespräche" und „Arzt-Patienten-Gespräche".

Die Gesprächsdaten aus der Datenbank *Gesprochenes Deutsch für die Auslandsgermanistik* werden von Auslandsgermanisten aus über 30 Ländern genutzt, und mittlerweile existieren mehrere Arbeiten, die auf der Grundlage der Daten die Relevanz der Vermittlung von Phänomenen der gesprochenen Sprache für den DaF-Unterricht herausstellen bzw. aus dem Datenmaterial konkrete Didaktisierungsvorschläge erarbeiten. So liegen beispielsweise Anregungen zur Vermittlung von Vergewisserungssignalen (Imo 2011), regionalen Varietäten (Imo 2012), progressiven *am*-Konstruktionen (Imo 2015), Funktionen von *ja* im Gespräch (Weidner 2015), zum Erwerb von Erzählanfängen (Zitta 2015), zur Modalpartikel *halt* (Betz 2015) und zu *weil* (Bendig, Betz & Huth 2016) vor.

Die Tatsachen, dass die Datenbank lediglich der Auslandsgermanistik zur Verfügung steht (und somit für DaZ nicht verwendet werden kann), dass die Anzahl der Gespräche und Gesprächstypen vergleichsweise gering ist, und

dass es sich nicht im eigentlichen Sinn um eine Datenbank, sondern eher um ein Archiv handelt, schränkt die Nutzbarkeit dieser Ressource allerdings ein. Um diese Mängel zu beheben, wurde im Jahr 2017 mit Förderung des Ministeriums für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen an der Universität Münster eine Datenbank aufgebaut, die verschiedene – speziell für die DaF- und DaZ-Vermittlung erhobene – Gesprächstypen beinhaltet.

4 Die *Plattform Gesprochenes Deutsch* – ein Korpus für Forschung und Praxis in DaF/DaZ

4.1 Vorstellung der Datenbank

Um eine Ressource zu schaffen, die neben der Auslandsgermanistik auch für die Inlandsgermanistik bzw. Sprachkurse in Deutschland von Nutzen ist, wird im Rahmen des Projekts *Plattform Gesprochenes Deutsch – authentische Alltagsinteraktionen für die Forschung und Praxis im Bereich DaF und DaZ* (http://dafdaz.sprache-interaktion.de/ [letzter Zugriff: 21. 1. 2017]) eine Internet-Plattform eingerichtet, die eine umfangreiche Sammlung von Gesprächen deutscher Muttersprachler zum Download bereitstellt. Die Interaktionssituationen entstammen unterschiedlichen informellen wie auch institutionellen Kontexten und enthalten verschiedene kommunikative Handlungen und Gesprächsgattungen. Die Gesprächsdaten sind darüber hinaus auch der DaF- und DaZ-Lehrerausbildung an Universitäten sowie der Forschung zugänglich. Somit können sie auch von Sprachwissenschaftlern, Deutschdidaktikern, Interaktionsforschern sowie von Pädagogen und Soziologen als Datenbasis für Forschungsprojekte genutzt werden.

Die im Aufbau befindliche Datensammlung umfasst bisher neben privaten Face-to-face-Gesprächen und Telefongesprächen im Familien- und Freundeskreis auch Interaktionen aus dem Hochschulkontext (Studienberatung, Beantragung eines Bibliotheksausweises, Seminarplanung, wissenschaftliche Vorträge), Verkaufsgespräche (u. a. in der Buchhandlung, an der Kinokasse, beim Bäcker), kurze Beratungsgespräche (in der Reinigung, beim Juwelier, beim Friseur), Bestellungen und Bezahlvorgänge in Restaurants und Tankstellen, Arzt-Patienten-Interaktionen und Gespräche in Behörden und Ämtern. Die Gesprächsausschnitte sind nach thematischen Gruppen geordnet und können als Audio- bzw. Videodateien sowie in Form von Transkripten heruntergeladen werden. Die verwendeten Transkriptionskonventionen nach GAT 2 (Selting et al.

2009) werden für Nutzer, die im Umgang mit Transkripten nicht geschult sind, auf den Seiten der Plattform erläutert. Mittels Suchfunktion können Lehrende in den Gesprächen gezielt bestimmte Phänomene recherchieren.

Basierend auf den Gesprächsdaten werden Lehreinheiten zu Strukturmerkmalen der gesprochenen Sprache (z. B. zum Sprecherwechsel, zu Paarsequenzen, zu Diskursmarkern, zu Herausstellungsstrukturen, zu Höflichkeit im Gespräch etc.) zur Verfügung gestellt, in denen Formen und Funktionen von Strukturmerkmalen der Mündlichkeit lernergerecht didaktisiert sind. So bietet die Plattform eine reichhaltige Datenbasis, um folgende essenzielle Kompetenzen der Deutschlernenden mit dem Lernziel Alltagssprache zu schulen:
- Hörverstehen,
- Gattungswissen typischer Kommunikationssituationen im Alltag,
- Wissen über interaktive Strukturen in Gesprächen,
- situationsangemessene Sprechstile,
- syntaktische Konstruktionen des Gesprochenen Deutsch,
- sprachliche Variation,
- phonotaktische und prosodische Muster im Gesprochenen Deutsch,
- Landeskunde,
- Wortschatz,
- körperlich-visuelle Kommunikationsressourcen.

Neben den Lehreinheiten, die als Worddokumente heruntergeladen werden können, sind derzeit auch E-Learning-Einheiten im Aufbau, die online absolviert werden können. Sie ermöglichen Lernern abseits institutionalisierter Bildungswege anhand von exemplarisch ausgewählten mündlichen Phänomenen den selbstständigen Erwerb mündlicher Sprachkompetenz und kulturspezifischen Wissens über alltägliche Interaktionssituationen. Auf diese Weise versucht die Plattform der sowohl in der Forschung als auch in der Praxis stets eingeforderten Vermittlung zwischen der sprachwissenschaftlichen Auseinandersetzung und Beschreibung authentischer Alltagssprache und deren Didaktisierung für die Praxis Rechnung zu tragen. Da das Korpus nicht auf den DaF- und DaZ-Lehrkontext beschränkt ist, können auch Studierende und Forschende die Daten als Grundlage zur Untersuchung interaktionaler Strukturen, syntaktischer Konstruktionen, prosodischer Muster sowie körperlich-visueller Kommunikationsressourcen nutzen. Die Erkenntnisse dieser Forschung können dann wiederum der Sprachdidaktik für die Lehre zur Verfügung gestellt werden.

4.2 Beispielanalysen mit Anwendungsbezug für den DaF- und DaZ-Unterricht

Anhand zweier exemplarisch ausgewählter Phänomene mündlicher Alltagssprache wird im Folgenden verdeutlicht, inwiefern die Arbeit mit authentischen Gesprächsausschnitten aus der Datenbank für den DaF-Unterricht, und ganz besonders für den DaZ-Unterricht, gewinnbringend ist. Dabei wird der Fokus auf die interaktive Konstitution von Gesprächen gelegt, die in vielen Lehrwerken nicht eigens zum Lerngegenstand erhoben wird.[3] Zum einen wird mit der Leistung von Höreraktivitäten zur Aufrechterhaltung der Kommunikation ein Mikrophänomen der Interaktion thematisiert, zum anderen wird mit der kommunikativen Handlung des Bestellens in einem Café ein komplexeres sequenzielles Muster diskutiert.

Untersuchungen von Lehrwerkdialogen (z. B. Bachmann-Stein 2013; Günthner, Wegner & Weidner 2013) zeigen, dass gesprächsorganisatorische Aktivitäten, wie Höreraktivitäten, in den konstruierten Hörtexten vernachlässigt werden. Ihr Fehlen kann in der authentischen Kommunikation Folgen haben. Besonders in Telefonaten würde die Kommunikation rasch zusammenbrechen, wenn der Hörer keine regelmäßigen Signale senden würde, die Aufmerksamkeit und/oder Verstehen dokumentieren. Anhand von authentischen Gesprächsausschnitten können Lernende für die sequenzielle Platzierung sowie für die formale und funktionale Variation von Höreraktivitäten sensibilisiert werden. Das Lernziel liegt also im Erwerb von Interaktionskompetenzen, die durch Lehrwerke bislang nicht zufriedenstellend vermittelt wurden. Vor diesem Hintergrund kann dem von Rösler (2016) zurecht vorgebrachten Einwand gegen den Einsatz von Transkripten begegnet werden:

> Wenn im Unterricht mit Transkripten gearbeitet werden soll, dann muss es eine didaktisch sinnvolle Begründung dafür geben. Die unbestreitbare Tatsache, dass sie gesprochene Sprache genauer wiedergeben, reicht allein als Begründung nicht aus. Die Arbeit mit einem Transkript in einer bestimmten Form muss für eine konkrete Gruppe von Lernenden auf ihrem Sprachniveau machbar sein und es muss eine Aufgabe/ein Lernziel geben, das den Mehraufwand rechtfertigt. (Rösler 2016: 145)

Höreraktivitäten sind in ihrer Form und Funktion vielfältig. In der Duden Grammatik (2009) werden *hmhm* und *ja* als die am häufigsten auftretenden Höreraktivitäten genannt, die in den meisten Fällen lediglich Aufmerksamkeit

---

3 Die Relevanz typischer syntaktischer Konstruktionen des Mündlichen für den Sprachunterricht mit Nicht-Muttersprachlern wurde bereits in mehreren Studien dargestellt. Vgl. exemplarisch Moraldo & Missaglia (2013), Schneider (2013), Imo (2015, 2016) und Weidner (i. E.).

signalisieren. Sie sind nicht klar zu trennen von Höreraktivitäten, die eine responsive Funktion haben (etwa *aha*, *klar*, *stimmt* etc.). Auch Interjektionen wie *boah*, *wow* o. ä. kommen als Höreraktivitäten vor und drücken Erstaunen oder Entsetzen aus. Die allen Höreraktivitäten gemeinsame Grundfunktion liegt darin, dass mit ihnen nicht die Sprecherrolle übernommen wird, sondern dass mit ihnen der Sprecher in seiner Rolle bestätigt wird. Der untenstehende Transkriptausschnitt ist der Datenbank *Plattform Gesprochenes Deutsch* entnommen. Inhaltlich geht es darum, dass Familienmitglieder über einen Zeitschriftenartikel sprechen, in dem eine Journalistin das staatliche Zeugenschutzprogramm kritisiert.

(1) Gespräch über Zeugenschutzprogramm

```
133 MR:   war_n MEHrere solche fälle.
134 LR:   [HM_hm; ]
135 CR:   [ja     ] aber INnerhalb deutschland,=
136       =es GIBT ja auch den fall,
137       dass man quasi n mh äh ameriKAnischen pass dann bekommt,
138 MR:   in DEM fall war_s innerhalb [deutschlands] ja;
139 CR:                               [oKAY;       ]
140 LR:   °h weil im fernsehen wird_s ja immer SO dargestellt,
141       als ob du rund um die [uhr] n SCHUTZ hättest;
142 CR:                         [ja;]
143 LR:   und dass die sich wirklich beMÜhen [gell;]
144 MR:                                      [bei  ] EIner,
145       hat_se DARgestellt,
146       stand sogar der ehemalige NAme am am klingelschild;
147 LR:   °ho:;
148 MR:   (-) also DES zu thema [schutz; ja,]
149 LR:                         [boa:h;     ]
```

Gesprächsteilnehmerin LR gibt mit dem Hörersignal *HM_hm;* (Z. 134) zu verstehen, dass sie MR zuhört und seiner Geschichte folgt. Das von CR geäußerte *oKAY;* (Z. 139) ist dagegen eine Höreraktivität, die eine responsive, aber auch kontinuierende Funktion hat. *Okay* quittiert die Information aus Zeile 138, die eine Bestätigung von CRs Vermutung (Z. 135) darstellt, der erzählte Fall habe innerhalb Deutschlands stattgefunden. Mit der Höreraktivität *ja* (Z. 142) stimmt CR der Aussage von LR über die Darstellung des Zeugenschutzprogramms im Fernsehen zu. Die Interjektionen in den Zeilen 147 und 149 stellen Höreraktivitäten dar, mit denen jeweils Stellung zum geäußerten Sachverhalt genommen

wird. Bereits die sequenzielle Platzierung nach der vollständig abgeschlossenen TCU (turn constructional unit) in Zeile 146 – und eben nicht in Überlappung – deutet darauf hin, dass es sich bei *°ho:;* nicht um ein Hörersignal im Sinne eines bloßen „Ich verstehe, mach weiter“ handelt. Mit Goffman (1978) sind die Höreraktivitäten als *Response Cries* zu werten und drücken Entsetzen über den in den Zeilen 145–147 geschilderten Fehler der Behörden aus. Eine Nicht-Reaktion wäre an dieser Stelle markiert, da der berichtete Sachverhalt im Kontext des Gesprächs eine ‚unerhörte Begebenheit‘ darstellt.

Anhand dieses Ausschnitts zeigt sich einerseits, dass die gleiche sprachliche Form einer Höreraktivität (etwa *ja* mit fallender Intonation) für verschiedene Zwecke eingesetzt werden kann, und andererseits, dass die gleiche Funktion (z. B. Entsetzen ausdrücken) durch unterschiedliche Formen (hörbares Einatmen durch *°ho* und *boah*) realisiert werden kann. Bereits Ehlich (1979) und auch Zifonun et al. (1997) stellen das weit gefächerte Formen- und Funktionsspektrum der Höreraktivität *hmhm* ausführlich dar. Imo (2013c: 292 ff.) und Weidner (2015) diskutieren Möglichkeiten, Deutschlernenden die sequenziellen, prosodischen und phonetischen Formen sowie die Funktionen von *ja* im Gespräch näherzubringen. Am Beispiel von Höreraktivitäten zeigt sich, dass die Fähigkeit, grammatisch korrekte Sätze zu bilden, nicht ausreicht, um als kompetenter Sprecher in mündlichen Situationen agieren zu können. Wenn in Gesprächen mit Nicht-Muttersprachlern Höreraktivitäten ausbleiben, eine unpassende Form oder eine prosodische Realisierung gewählt wird, die in der sequenziellen und inhaltlichen Umgebung inadäquat wirkt, kann dies zu Irritationen führen, etwa im Hinblick darauf, ob das Gesagte verstanden oder etwa als unhöfliches Verhalten wahrgenommen wurde.

Nicht nur am Beispiel von Partikeln kann mittels authentischer Dialoge für interaktionale Strukturen sensibilisiert werden. Auch komplexere Gesprächsstrukturen können zum Thema gemacht werden. Sowohl der DaF- als auch der DaZ-Unterricht verfolgen mit ihrem handlungsorientierten Ansatz das Ziel, Lernende in die Lage zu versetzen, die kommunikativen Aufgaben des Alltags kompetent zu bewältigen. Das Lernpotenzial der Gespräche in der *Plattform Gesprochenes Deutsch* liegt u. a. darin, dass ihnen – im Gegensatz zu den konstruierten Lehrwerkdialogen – eine ‚echte‘ Handlungsqualität zukommt. Durch die Beschäftigung mit Verfahren zur Bewältigung spezifischer kommunikativer Aufgaben können Lernenden „realitätstaugliche [...] sprachliche [...] Handlungsmuster an die Hand gegeben [werden], die außerhalb des Unterrichts einsetzbar“ (Bachmann-Stein 2013: 43) sind. Beispielhaft kann dies am folgenden kurzen Transkriptausschnitt einer Bestellung im Café gezeigt werden, in dem der Gast (CR) bei der Bedienung (FT) einen Espresso und ein Leitungswasser bestellt:

(2) Bestellen im Café

```
Der Gast (CR) bestellt im Café bei der Kellnerin (FT) ein Getränk. Das
Café wird hauptsächlich von Studierenden frequentiert; es herrscht eine
entspannte Stimmung mit Hintergrundmusik.

003 CR:   HALlo;                                  } Gruß
004 FT:   HI_i,                                   } Gruß

005       darf_s schon was für dich SEIN?         } Aufforderung
006 CR:   äh ja ich nehm einen esPRESso und       } Nachkommen
          ein leitungswasser bitte.
007 FT:   GERne.                                  } Ratifizierung

008       BRING ich dir;                          } Ankündigung
009 CR:   DANke;                                  } Dank
```

Dass Gesprächseinstiege hochgradig strukturiert ablaufen, zeigte bereits Schegloff (1968) anhand von Telefonanfängen. Auch die Bestellung in einem Café oder Restaurant folgt einem routinierten Muster, das durch Paarsequenzen strukturiert ist. In den Zeilen 003 und 004 des Beispiels in (2) findet sich eine Gruß-Gegengruß-Sequenz, hier in Minimalform durch *Hallo* und *Hi* realisiert. Die Bedienung schließt mit einem weiteren ersten Teil einer Paarsequenz an, nämlich der Aufforderung an den Gast, einen Bestellwunsch zu äußern. Diese Aufforderung bringt sie höflich in Form einer Frage (Z. 005) vor. Der konditionell relevant gesetzte zweite Paarteil besteht zunächst darin, dass der Gast die ihm gestellte Frage bejaht, um dann die konkrete Bestellung zu nennen (Z. 006). Auch die Bestellhandlung macht wiederum einen Redezug erwartbar, und zwar die Ratifizierung durch die Bedienung, die hier durch *gerne* (Z. 007) erfolgt. Das Muster besteht demnach aus dem Dreischritt „Aufforderung, Nachkommen der Aufforderung, Ratifizierung". Die Bedienung schließt die Bestellung mit der Ankündigung ab, die Bestellung umzusetzen (*bring ich dir*, Z. 008). Diese Ankündigung der Serviceleistung macht schließlich den Dank des Gastes relevant, der daraufhin erfolgt (Z. 009) und das Ende der Bestellaktivität markiert.

Neben dem Einüben der einzelnen formelhaften Paarteile, aus denen sich die sequenzielle Struktur von Bestellaktivitäten (Begrüßung, Bestellung, Dank) zusammensetzt, kann im Unterricht zudem das jeweils Kulturspezifische der Aktivität herausgearbeitet werden. Beispielweise ist zu erwähnen, dass die Gäste in deutschen Cafés warten, bis die Bedienung kommt, um die Bestellung aufzunehmen, und dass es unüblich ist, laut durch das Café nach der Bedienung zu rufen. Auch kann anhand des Ausschnitts die Situationsangemessenheit des Sprechens thematisiert werden. In Bezug auf die Begrüßung ist hierfür die im Transkriptkopf (d. h. in der über dem Transkript befindlichen Erläuterung) erwähnte Information wichtig, dass es sich im Beispiel um ein Café mit

hauptsächlich studentischem Publikum handelt. In dem Zusammenhang können verschiedene Begrüßungsformeln in ihrer Angemessenheit in verschiedenen Kontexten diskutiert werden (etwa die Frage, wie eine Begrüßung in einem konservativen oder gehobenen Restaurant aussähe, etc.). Mit Blick auf die Aufforderung zum Bestellen (Z. 005) sowie die konkrete Äußerung der Bestellung (Z. 006) könnten alternative syntaktische Formate, die im Deutschen formelhaft verfestigt sind (*Was darf es sein?; Haben Sie bereits gewählt?* und *Ich hätte gern ...; Für mich bitte ...*), gesammelt werden. Im Zuge dessen bietet es sich auch an, auf die Möglichkeiten der Markierung von Höflichkeit, etwa durch die Verwendung von Modalverben, des Konjunktivs und Höflichkeitspartikeln wie *bitte* einzugehen. Die derzeit entwickelten Didaktisierungsvorschläge, die auf den Internetseiten der *Plattform Gesprochenes Deutsch* zum Download bereitgestellt werden, geben Lehrkräften Hilfestellungen, um Phänomene wie die oben vorgestellten im DaF- oder DaZ-Unterricht zu behandeln. Daneben werden Literaturhinweise gegeben, die Lehrenden den Zugang zu Studien der Erforschung gesprochener Sprache erleichtern, um so die Vermittlung von Forschungsergebnissen in die Praxis voranzutreiben.

5 Fazit

Das Ziel des vorliegenden Beitrags bestand darin, dafür zu plädieren, mündliche Korpora als Datenquelle im Fremdsprachenunterricht einzusetzen. Es liegt natürlich auf der Hand, dass die Verwendung von authentischen Alltagsinteraktionen für die Lernenden deutlich komplexer – und für die Lehrenden entsprechend zeitaufwändiger – ist. Der vorliegende Beitrag soll daher nicht als ein Plädoyer für den *ausschließlichen* Einsatz von authentischen Daten verstanden werden. Uns ist bewusst, dass eine Fokussierung auf einen Themenbereich beinahe zwangsläufig zu einer Reduktion anderer Bereiche führen wird, wie Hirschfeld, Rösler & Schramm (2016: 134) in ihrer Diskussion von „Mündlichkeit" im Fremdsprachenunterricht feststellen:

> Eine Erweiterung der Gegenstände, die beim gesteuerten Fremdsprachenlernen im Unterricht behandelt werden, geht nicht automatisch einher mit mehr Zeit für das Fremdsprachenlernen in Bildungsinstitutionen. Dies führt dazu, dass eine Diskussion über Auswahlentscheidungen für die Aufnahme von Gegenständen in das Curriculum geführt werden muss. (Hirschfeld, Rösler & Schramm 2016: 134)

In welchem Umfang und an welchen Stellen im Unterricht authentische mündliche Interaktionen eingesetzt werden sollen, muss daher von den Lehrerenden jeweils in Abstimmung auf Lehrpläne und Interessen und Fähigkeiten der

Lernergruppen entschieden werden. Dabei geht es „nicht um gesprochene vs. geschriebene Sprache“ im Sinne eines Ausspielens beider Bereiche gegeneinander, sondern lediglich um die ziel- und situationsorientierte didaktische Frage, „was warum wann wie lange und wie oft auf welche Weise für wen“ (Rösler 2016: 137) vermittelt wird. Im Kontext des DaF-Unterrichts bestimmen dabei Aspekte wie die Nähe zu Deutschland (und entsprechend die Wahrscheinlichkeit von Aufenthalten dort), bevorstehende Studienaufenthalte in Deutschland, Vorgaben der DaF-Lehrpläne und die Ziele der Studierenden (geht es um einen Vorbereitungskurs für ein Philosophiestudium, in dem deutsche Texte gelesen werden sollen, um einen Fokus auf schriftliches Übersetzen, eine Ausbildung zum Simultandolmetscher oder um den Erwerb von umfassenden – auch mündlichen – Kompetenzen, um z. B. Deutschlehrer zu werden?) den möglichen Anteil authentischer Mündlichkeit. Ganz anders sieht es im Kontext des DaZ-Unterrichts aus. Hier kommt der mündlichen Kompetenz (ganz besonders bei Kursen für Flüchtlinge) eine sehr große Rolle zu, da der Unterricht zu einer Teilnahme am kommunikativen Alltag der Zielgesellschaft befähigen soll. Entsprechend sollten im DaZ-Bereich sowohl im Unterricht als auch im angeleiteten Selbststudium authentische Gesprächsdaten in hohem Umfang zum Einsatz kommen. Die hier vorgestellte Datenbank gibt sowohl den Lehrenden als auch den Lernenden im Selbststudium dafür ein einfach zu bedienendes Instrument an die Hand.

Literatur

Bachmann-Stein, Andrea (2013): Authentische Gesprochene Sprache im DaF-Unterricht: Pro und Contra. In Sandro M. Moraldo & Federica Missaglia (Hrsg.), *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*, 39–58. Heidelberg: Winter.

Bartz, Thomas (2015): Digitale Sprachressourcen im Deutschunterricht: Korpus-basierte Recherche und Analyse in der ‚Wörterbuchwerkstatt‘. In Wolf-Dirk Skiba & Alessandra Lombardi (Hrsg.), *Korpora im Sprachunterricht*, 1–8. Bozen: Bozen-Bolzano University Press.

Bartz, Thomas & Nadja Radtke (2014): Digitale Korpora im Deutschunterricht: Didaktisches Potenzial. *Zeitschrift für germanistische Linguistik* 42, 130–143.

Becker-Mrotzek, Michael & Gisela Brünner (2006): *Gesprächsanalyse und Gesprächsführung*. Radolfzell: Verlag für Gesprächsforschung.

Bendig, Ina, Emma Betz & Thorsten Huth (2016): “weil – das ist eben doch richtig so.” Teaching variant types of *weil-* and *obwohl*-structures in German. *Unterrichtspraxis/ Teaching German* 49, 214–227.

Betz, Emma (2015): “des is halt so”: Explaining, justifying, and convincing with ‘halt’. *Unterrichtspraxis/Teaching German* 48, 114–132.

Civegna, Klaus (2005): Authentische Texte. In Hans-Ludwig Krechel (Hrsg.), *Mehrsprachiger Fachunterricht in Ländern Europas*, 168–172 Tübingen: Narr.

Duden Grammatik (2009): *Duden: die Grammatik* (Duden 4). Hrsg. von der Dudenredaktion. 8., überarb. Aufl. Mannheim u. a.: Dudenverlag.
Edelhoff, Christoph (Hrsg.) (1985): Authentizität im Fremdsprachenunterricht. In Christoph Edelhoff (Hrsg.), *Authentische Texte im Deutschunterricht*, 7–30. Ismaning: Hueber.
Ehlich, Konrad (1979): Formen und Funktionen von ‚HM' – eine phonologisch-pragmatisch Analyse. In Harald Weydt (Hrsg.), *Die Partikeln der deutschen Sprache*, 503–517. Berlin, New York: de Gruyter.
Europarat (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Berlin: Langenscheidt.
Fandrych, Christian & Erwin Tschirner (2007): Korpuslinguistik und Deutsch als Fremdsprache. *Deutsch als Fremdsprache* 44, 195–204.
Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (2014): Das Korpusprojekt „Gesprochene Wissenschaftssprache kontrastiv" und seine Relevanz für die Vermittlung des Deutschen als Wissenschaftssprache. In Nicole Mackus & Jupp Möhring (Hrsg.), *Wege für Bildung, Beruf und Gesellschaft – mit Deutsch als Fremd- und Zweitsprache*, 141–160. Göttingen: Göttinger Universitätsverlag.
Feilke, Helmuth (2012): Transitorische Normen. In Susanne Günthner, Wolfgang Imo, Jan Georg Schneider & Dorothee Meer (Hrsg.), *Kommunikation und Öffentlichkeit*, 153–182. Berlin, Boston: de Gruyter.
Freie und Hansestadt Hamburg (2011): Bildungsplan Stadtteilschule Jahrgangsstufen 5–11, Deutsch als Zweitsprache in Vorbereitungsklassen. Hamburg: Behörde für Schule und Bildung.
Goffman, Erving (1978): Response cries. *Language* 54, 787–815.
Günthner, Susanne (2000): Grammatik der gesprochenen Sprache – eine Herausforderung für Deutsch als Fremdsprache? *Info DaF* 27, 352–366.
Günthner, Susanne (2011): Übergänge zwischen Standard und Non-Standard – welches Deutsch vermitteln wir im DaF-Unterricht? In Eva L. Wyss & Daniel Stotz (Hrsg.), *Sprachkompetenz in Ausbildung und Beruf*, 24–47. Neuenburg: VALS-ASLA.
Günthner, Susanne, Lars Wegner & Beate Weidner (2013): Gesprochene Sprache im DaF-Unterricht. In Sandro M. Moraldo & Federica Missaglia (Hrsg.), *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*, 113–150. Heidelberg: Winter.
Gut, Ulrike (2007): Sprachkorpora im Phonetikunterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 13, 1–21.
Gut, Ulrike (2009): *Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt am Main: Peter Lang.
Handwerker, Brigitte & Karin Madlener (2009): *Chunks für DaF. Theoretischer Hintergrund und Prototyp einer multimodalen Lernumgebung*. Baltmannsweiler: Schneider Hohengehren.
Handwerker, Brigitte, Rainer Bäuerle & Bernd Sieberg (Hrsg.) (2016): *Gesprochene Fremdsprache Deutsch*. Baltmannsweiler: Schneider Hohengehren.
Hennig, Mathilde (2002): Wie kommt die gesprochene Sprache in die Grammatik? *Deutsche Sprache* 30, 307–326.
Hirschfeld, Ursula, Dieter Rösler & Karen Schramm (2016): Facetten der Mündlichkeit im DaF-Unterricht. Zur Einführung in den Themenschwerpunkt. *Deutsch als Fremdsprache* 3, 131–134.
Imo, Wolfgang (2009): Welchen Stellenwert sollen und können Ergebnisse der Gesprochene-Sprache-Forschung für den DaF-Unterricht haben? In Andrea Bachmann-Stein &

Stephan Stein (Hrsg.), *Mediale Varietäten: Analysen von gesprochener und geschriebener Sprache und ihre fremdsprachendidaktischen Potenziale*, 39–61. Landau: VEP.

Imo, Wolfgang (2011): ‚Jetzt gehn wir einen trinken, gell?' Vergewisserungssignale (*tag questions*) und ihre Relevanz für den DaF-Unterricht. In Sandro M. Moraldo (Hrsg.), *Deutsch aktuell 2. Einführung in die Tendenzen der Gegenwartssprache*, 127–150. Rom: Carocci.

Imo, Wolfgang (2012): Hattu Möhrchen? Gesprochene Sprache im DaF-Unterricht. In Ulrike Reeg, Pasquale Gallo & Sandro Moraldo (Hrsg.): *Gesprochene Sprache im DaF-Unterricht. Zur Theorie und Praxis eines Lerngegenstandes*, 29–56. Münster: Waxmann.

Imo, Wolfgang (2013a): ‚Rede' und ‚Schreibe': Warum es sinnvoll ist, im DaF-Unterricht beides zu vermitteln. In Sandro M. Moraldo & Federico Missaglia (Hrsg.): *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*, 55–78. Heidelberg: Winter.

Imo, Wolfgang (2013b): Authentisches gesprochenes Deutsch im DaF-Unterricht. In Cordula Schulze (Hrsg.), *Die deutsche Sprache und ich*, 49–62. Münster: LIT.

Imo, Wolfgang (2013c*): Sprache in Interaktion: Analysemethoden und Untersuchungsfelder*. Berlin, Boston: de Gruyter Mouton.

Imo, Wolfgang (2015): Aspektrealisierung im gesprochenen Deutsch zwischen Norm und Gebrauch. In Sandro M. Moraldo & Wolfgang Imo (Hrsg.), *Interaktionale Sprache im DaF-Unterricht*, 367–393. Tübingen: Stauffenburg.

Imo, Wolfgang (2016): Ich finde, mit Matrixsätzen kann man eine Menge machen … Von der Redeanführung über den Matrixsatz zum Diskursmarker. In Brigitte Handwerker, Rainer Bäuerle & Bernd Sieberg (Hrsg.), *Gesprochene Fremdsprache Deutsch*, 45–74. Baltmannsweiler: Schneider Hohengehren.

Imo, Wolfgang (i. V.): Der Einsatz authentischer Alltagssprachdaten im DaF-Unterricht: Theorie und Praxis. In Song Chol Park (Hrsg.), *Jahrbuch für Internationale Germanistik*.

Imo, Wolfgang & Sandro M. Moraldo (Hrsg.) (2015): *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*. Tübingen: Stauffenburg.

Lemnitzer, Lothar & Heike Zinsmeister (2015): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Liedtke, Martina (2013): Mit Transkripten Deutsch lernen. In Sandro M. Moraldo & Federica Missaglia (Hrsg.), *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*, 243–266. Heidelberg: Winter.

Lüdeling, Anke & Maik Walter (2010): Korpuslinguistik. In Hans-Jürgen Krumm, Christian Fandrych, Britta Hufeisen & Claudia Riemer (Hrsg.), *Handbuch Deutsch als Fremd- und Zweitsprache*, 1. Halbband (Handbücher zur Sprach- und Kommunikationswissenschaft 35.1), 315–322. Berlin, Boston: de Gruyter Mouton.

Lüger, Heinz-Helmut (2009): Authentische Mündlichkeit im fremdsprachlichen Unterricht? *Beiträge zur Fremdsprachenvermittlung* 15, 15–37.

Mac, Agnieszka (2011): Zum Einsatz von authentischem Quellenmaterial im Fremdsprachenunterricht am Beispiel von Fernsehnachrichten. *Glottodidactica* XXXVII, 73–84.

Moraldo, Sandro M. (Hrsg.) (2011): *Deutsch aktuell 2. Einführung in die Tendenzen der deutschen Gegenwartssprache*. Rom: Carocci.

Moraldo, Sandro M. & Federica Missaglia (Hrsg.) (2013): *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*. Heidelberg: Winter.

Pieklarz, Magdalena (2010): Gesprochene Sprache in der philologischen Sprachausbildung – Grenzen und wechselnde Herausforderungen. In Natalia S. Babenko & Natalia A. Bakshi

(Hrsg.), *Russische Germanistik. Jahrbuch des Russischen Germanistenverbandes*, 257–273. Moskau: DAAD.
Pieklarz-Thien, Magdalena (2015): *Gesprochene Sprache in der philologischen Sprachausbildung*. Frankfurt am Main: Lang.
Reeg, Ulrike, Pasquale Gallo & Sandro M. Moraldo (Hrsg.) (2012): *Gesprochene Sprache im DaF-Unterricht. Zur Theorie und Praxis eines Lerngegenstandes*. Münster: Waxmann.
Reershemius, Gertrud & Evelyn Ziegler (2015): Sprachkontaktinduzierte jugendkulturelle Stile im DaF-Unterricht: Beispiele aus dem Film „Fack ju Göhte". In Wolfgang Imo & Sandro M. Moraldo (Hrsg.), *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*, 234–278. Tübingen: Stauffenburg.
Rösler, Dieter (2016): Nähe und Distanz zur Mündlichkeit in der fremdsprachendidaktischen Diskussion. *Deutsch als Fremdsprache* 3, 135–148.
Sächsisches Staatsministerium für Kultus (2009): Lehrplan für Vorbereitungsgruppen, Vorbereitungsklassen, Vorbereitungsklassen mit berufspraktischen Aspekten Deutsch als Zweitsprache. Dresden: Sächsisches Staatsministerium für Kultus.
Schegloff, Emanuel A. (1968): Sequencing in Conversational Openings. *American Anthropologist* 70, 1075–1095.
Schneider, Jan Georg (2013): ‚die war letztes mal (–) war die länger' – Überlegungen zur linguistischen Kategorie ‚gesprochenes Standarddeutsch' und ihre Relevanz für die DaF-Didaktik. In Sandro M. Moraldo & Federica Missaglia (Hrsg.), *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis*, 83–111. Winter: Heidelberg.
Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock & Susanne Uhmann (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, 353–402.
Sieberg, Bernd (2013): *Sprechen lehren, lernen und verstehen*. Tübingen: Stauffenburg.
Sieberg, Bernd (2016): Reaktive: Vorschlag für eine Erweiterung der Kategorie Responsive. In Brigitte Handwerker, Rainer Bäuerle & Bernd Sieberg (Hrsg.), G*esprochene Fremdsprache Deutsch*, 101–116. Baltmannsweiler: Schneider Hohengehren.
Siepmann, Dirk (2009): Korpuslinguistik und Fremdsprachenunterricht. In Udo O. H. Jung (Hrsg.), *Praktische Handreichung für Fremdsprachenlehrer*, 321–330. Frankfurt am Main: Lang.
Sinclair, John (2004): *How to use corpora in language teaching*. Amsterdam: Benjamins.
Trouvain, Jürgen & Frank Zimmerer (2016): Phonetische Lernerkorpora und ihr Nutzen im DaF-Bereich – eine Fallstudie. *Deutsch als Fremdsprache* 4/2016, 2014–2021.
Wallner, Franziska (2013): Korpora im DaF-Unterricht. Potentiale und Perspektiven am Beispiel des DWDS. *Revista Nebrija de Linguistica Aplicada* 13, o.S.
Weidner, Beate (2012): Gesprochenes Deutsch für die Auslandsgermanistik – Eine Projektvorstellung. *Info DaF* 39, 31–51.
Weidner, Beate (2015): Das funktionale Spektrum von *ja* im Gespräch – ein Didaktisierungsvorschlag für den DaF-Unterricht. In Wolfgang Imo & Sandro M. Moraldo (Hrsg.), *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht*, 165–195. Tübingen: Stauffenburg.
Weidner, Beate (2018): Gesprochene Sprache als Unterrichtsgegenstand. In Zeynep Kalkavan-Aydin (Hrsg.), *DaZ/DaF-Didaktik – Praxishandbuch für die Sek. I und II*, 152–170. Berlin: Cornelsen.

Wittig, Matthias (2015): Jenseits von Vokabeln und Grammatik. In Japanische Gesellschaft für Germanistik (Hrsg.), *Mündliche Kommunikation im DaF-Unterricht. Phonetik, Gespräch und Rhetorik*, 149–169. München: iudicium.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker (1997): *Grammatik der deutschen Sprache*. 3 Bde. Berlin, New York: de Gruyter.

Zitta, Eva (2015): ‚weißte WAS?' – Erzählen im DaF-Unterricht. Zur Vermittlung gesprochensprachlicher Kompetenz im DaF-Unterricht am Beispiel von Erzähleinstiegen. In Sandro M. Moraldo & Wolfgang Imo (Hrsg.), *Interaktionale Sprache im DaF-Unterricht*, 113–131. Tübingen: Stauffenburg.

Register

https://doi.org/10.1515/9783110538649-012

Autorinnen und Autoren

Hans C. Boas, Department of Germanic Studies, The University of Texas at Austin, 1 University Station, C3300, 2505 University Blvd., Austin, Texas 78712, U.S.A., E-Mail: hcb@mail.utexas.edu

Noah Bubenhofer, Institut für Computerlinguistik, Universität Zürich, Andreasstr. 15, CH-8050 Zürich, Schweiz, E-Mail: bubenhofer@cl.uzh.ch

Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, E-Mail: dipper@linguistics.rub.de

Christoph Draxler, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: draxler@phonetik.uni-muenchen.de

Matthias Fingerhuth, Institut für Germanistik, Universität Wien, Universitätsring 1, A-1010 Wien, Österreich, E-Mail: matthias.fingerhuth@univie.ac.at

Stefan Th. Gries, Department of Linguistics, University of California Santa Barbara, Santa Barbara, CA 93106-3100; U.S.A., E-Mail: stgries@linguistics.ucsb.edu

Wolfgang Imo, Universität Hamburg, Institut für Germanistik, Überseering 35, D-22297 Hamburg, E-Mail: Wolfgang.Imo@uni-hamburg.de

Roland Kehrein, Forschungszentrum Deutscher Sprachatlas, Pilgrimstein 16, D-35032 Marburg, E-Mail: kehrein@uni-marburg.de

Marc Kupietz, Institut für Deutsche Sprache, R 5, 6–13, D-68161 Mannheim, E-Mail: kupietz@ids-mannheim.de

Sarah Kwekkeboom, Germanistisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, D-44801 Bochum, E-Mail: sarah.kwekkeboom@rub.de

Christian Mair, Englisches Seminar, Universität Freiburg, 79085 Freiburg, E-Mail: christian.mair@anglistik.uni-freiburg.de

Joachim Scharloth, Waseda University, School of International Liberal Studies, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050 Japan, E-Mail: scharloth@waseda.jp

Florian Schiel, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: schiel@phonetik.uni-muenchen.de

Thomas Schmidt, Institut für Deutsche Sprache, R 5, 6–13, Büro 3.13, D-68161 Mannheim, E-Mail: thomas.schmidt@ids-mannheim.de

Lars Vorberger, Forschungszentrum Deutscher Sprachatlas, Pilgrimstein 16, D-35032 Marburg, E-Mail: lars.vorberger@deutscher-sprachatlas.de

Beate Weidner, Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6–8, D-45127 Essen, E-Mail: Beate.Weidner@uni-due.de

https://doi.org/10.1515/9783110538649-013